Committee on National Statistics

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

BEA Expert Meeting on Exploiting Commercial Data for Official Economic Statistics

Thursday, November 19, 2015

The NAS Building

Board Room

Washington, DC

CONTENTS

BEA Expert Meeting on Exploiting Commercial Data for Official Economic Statistics

Thursday, November 19, 2015

The Board Room
The NAS Building

2101 Constitution Ave. NW
Washington, DC

AGENDA

<u>Meeting Objective</u>

Building on the recommendation in the Department of Commerce strategic plan to "explore, research, and test the extent to which government and private data can be shared," BEA asked CNSTAT to organize an expert meeting to explore the potential value added of (and obstacles to) using credit card company, retail sales, and other commercial information to improve the national accounts. The BEA requirement for expert guidance in these areas fits into the Department's broader visions for the use of big data to increase the quality of economic statistics, develop new statistical products, enhance the detail and timeliness of existing statistical products, and reduce costs and respondent burden.

This meeting will be oriented toward looking at specific applications, but should inform the broader effort by the federal statistical system to make effective use of alternative (non-survey) data sources. The main focus of this meeting will be on private sector transactions data. The meeting will be used as a venue for discussing next steps for further development of the data infrastructure of BEA.

9:00AM **Introductions**. ***John Abowd*** (Cornell), Moderator

(***Breakfast*** available outside the Board Room)

9:15 Current use of commercial data by BEA/Intro to issues—overview of commercial data sources drawn upon in the construction of the NIPAs. How are these data used to create more timely, detailed, or accurate estimates? ***Brian Moyer, David Johnson*** (BEA)

9:30 Overview: Potential big data applications. Challenges, opportunities and limitations of how alternative data sources (commercial, administrative) might be used to complement (or in some cases substitute) for survey collections in the construction of the NIPAs and in the production of official statistics more broadly. What are the methodological implications?

Trivellore Raghunathan (University of Michigan) – Statistical Challenges in Combining Information from Big and Small Data Sources

Abstract
Social Media, electronic transactional and administrative data, web scraping, and numerous other ways of collecting information have changed the landscape for those interested in addressing policy-relevant research questions. During the same time, the traditional sources of data, such as large-scale surveys, that have been a stable source for policy-relevant research have suffered setbacks due to large nonresponse and increasing cost of collecting such survey data. The non-survey data usually contain detailed information on certain behaviors on a large number of individuals (such as all credit card transactions) but very little background information on them (such as important covariates to address the policy-relevant question). On the other hand, survey data contains detailed information on the covariates but not so detailed information on the behaviors. Both data sources may not be representative of the target population of interest. This paper develops and evaluates a framework for linking information from multiple imperfect data sources along with the Census data to draw statistical inference. An explicit modeling framework involving selection into the big data, sampling and nonresponse mechanism in the survey data, distribution of the key variables of interest and certain marginal distributions from the Census Data are used as building blocks to draw inference about the population quantity of interest.

Simon Wilkie (Microsoft) Potential uses of commercial data in economic statistics. Possibilities of academic/public/private collaboration?

Discussant questions:

- What is the potential of expanded commercial data application—e.g., credit card and mobile payment information, data on retail sales, auto registrations, medical claims, and real estate sales, etc.—in further improving the accuracy, timeliness, and detail of the NIPAs? For reducing costs and respondent burden? For producing estimates at more granular levels of aggregation?

- Which NIPA components—e.g., spending patterns for various consumer retail categories, energy and other utility consumption, medical expenditures, housing, financial transactions—lend themselves to improved measurement from use of non-survey data sources?
- Models for combining data from multiple sources and evaluating the quality of estimates from them.

– ***Mick Couper*** (Michigan)
– ***John Haltiwanger*** (Maryland)
– ***Dennis Fixler*** (BEA)
– Open Discussion

11:00 ***Break***; refreshments available outside the Board Room

11:15 Presentations from private sector about kinds of data that they collect or generate, and how they are or may potentially be useful in construction of economic statistics. What is the evidence on how various estimates differ by data source—e.g., how well does credit card or mobile payment data track final consumption expenditures (at various levels of aggregation), or compare with Census Bureau retail data on spending from one category of good/service to another? (about 10 minutes each)

– JP Morgan Chase Institute (***Fiona Greig***)—big data for measuring financial activities and spending behaviors of individuals
– MasterCard (***Steve Tae, Kamalesh Rao***)
– Google (***Chris DiBona, Jeffrey Oldham***)—statistical uses of internal data (e.g., Google Consumer Survey, Google Survey Amplification) and computing issues for scaling big business data applications
– Zillow (***Stan Humphries***)—consumption of public record data; production and dissemination of derivative data based on raw input data
– Palantir/FirstData (**Alex Bores**)—visualizing relationships among, large amounts of data for the purpose of producing economic statistics
– Open discussion

12:30PM ***Lunch***; available outside the Board Room

1:30 Applications and issues with using commercial data in research

Jonathan Levin—exploiting commercial data for economic research, with extensions to official statistics, and specifically for improving the accuracy, timeliness, and detail of the GDP/national accounts

Matthew Shapiro—use of administrative account data for measuring spending, income, and assets

– Discussants: ***Katharine Abraham*** (Maryland/BEA Advisory Committee), ***Roberto Rigobon*** (MIT)

2:30 Discussion topics (and leaders):

– Big data projects at statistical agencies abroad—***Piet Daas*** (Statistics Netherlands)
– Using new technology to examine data, aggregate billions of data points. ***Simon Wilkie*** (Microsoft)
– How statistical thinking can help tackle the many Big Data challenges relevant to development of economic statistics. Scientific challenges facing broad disciplinary areas being transformed by Big Data and how statistical advances made in collaboration with other scientists can address these challenges. ***Frauke Kreuter*** (Maryland), ***Daniel Goroff*** (Alfred P. Sloan Foundation)
– Discussion of Privacy issues associated with blended government commercial data sources, ***Jerry Reiter*** (Duke)
– Public, Private data collaborations. How can synergies be exploited, and what are the incentives for collaboration. What are the complementarities (e.g., increased detail, timeliness and possibly reduced needed frequency of surveys in exchange for benchmarking, validating, quality control (trusted 3rd party)). How might additional cooperative arrangements be negotiated? What are the practical issues in obtaining data? ***Christopher Carroll*** (Consumer Federal Protection Bureau)

3:00 Refreshments available outside the Board Room

4:00 Conclude/Next Steps (***John Abowd, David Johnson***)

Role for CNSTAT? Coordination with the CNSTAT panel on a Multiple Data Sources Paradigm for Federal Statistics (Bob Groves, chair)? Agency follow-up and messages for the Federal Committee on Statistical Methodology working group?

4:15PM **Planned adjournment**

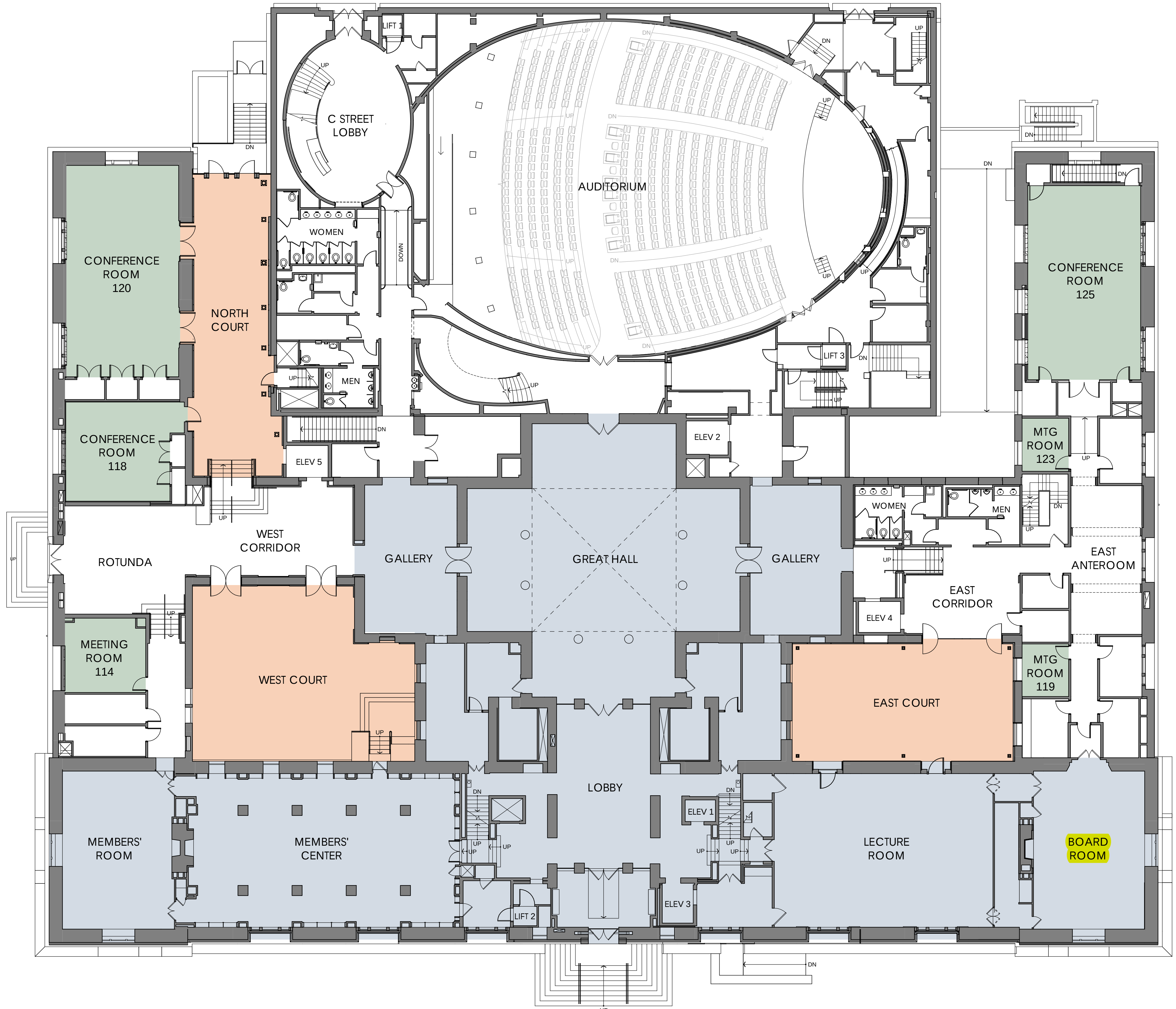
Summary Sheet

On entering the NAS Building's main lobby (C Street entrance), inform the guard at the desk that you are attending the ***BEA Commercial Data*** expert meeting in The Board Room, and present a photo ID.

Travel code	**DBASSE150143** Daily per diem rates: $179/night (hotel); $69/day (M&IE)
Agency	**Kentlands** (nas@uniglobekentlands.com; 1 800-552-6425)
Dates, times, location	**Thursday, November 19, 2015** (9:00 AM – 4:15 PM) The NAS Building, The Board Room (see attached floor plan) 2101 Constitution Avenue, NW Washington, DC 20037
Hotel details	**The State Plaza** 2117 E Street, NW, Washington, DC 20037 (**directions to NAS**)—C Street entrance
Meals and breaks	**Thursday, November 19** *Breakfast*: from 8:00 AM outside The Board Room *Break*: 11:00 AM refreshments available outside The Board Room *Lunch*: 12:30 PM available outside The Board Room
Complimentary parking	NAS garage on the corner of 20th Street NW and C Street NW
Metro	Foggy Bottom (Orange, Silver, and Blue lines)
A/V & communications	Data/video projector and screen, overhead projector Desktop with printer and Internet connection Free wireless Internet access (select Visitor; no password required)

Contacts

Chris Mackie: cmackie@nas.edu**;** Office: 202-334-2099
Anthony Mann: amann@nas.edu; Office: 202-334-3266; Cell: 703-472-8154



LIFT 1
C STREET LOBBY
AUDITORIUM
WOMEN
MEN
CONFERENCE ROOM 120
NORTH COURT
CONFERENCE ROOM 118
ELEV 5
ELEV 2
LIFT 3
CONFERENCE ROOM 125
MTG ROOM 123
ROTUNDA
WEST CORRIDOR
GALLERY
GREAT HALL
GALLERY
WOMEN
MEN
EAST ANTEROOM
EAST CORRIDOR
ELEV 4
MEETING ROOM 114
WEST COURT
EAST COURT
MTG ROOM 119
LOBBY
ELEV 1
MEMBERS' ROOM
MEMBERS' CENTER
LECTURE ROOM
BOARD ROOM
LIFT 2
ELEV 3

PARTICIPANT LIST

John Abowd, *ILR School, Cornell University* (Moderator)
Katharine Abraham, *MPRC, University of Maryland*
Alex Bores, *Palantir Technologies*
Bill Bostic, *U.S. Census Bureau*
Christopher Carroll, *Consumer Financial Protection Bureau*
Mick Couper, *PSC, University of Michigan*
Piet Daas, *Statistics Netherlands*
Chris DiBona, *Google, Inc.*
Austin Durrer, *ESA*
John Eltinge, *BLS*
Denis Fixler, *BEA*
David Friedman, *BLS*
Ophir Frieder, *Georgetown University*
Mahsa Gholizadeh, *BEA*
Daniel Goroff, *Alfred P. Sloan Foundation*
Fiona Greig, *JPMorgan Chase & Co.*
Alexis Grimm, *BEA*
Anne Hall, *BEA*
John Haltiwanger, *Department of Economics, University of Maryland*
Carma Hogue, *U.S. Census Bureau*
Mike Horrigan, *BLS*
Ellen Hughes-Cromwick, *ESA*
Stan Humphries, *Zillow*
Ron Jarmin, *U.S. Census Bureau*
David Johnson, *BEA*
Frauke Kreuter, *MPRC, University of Maryland*
Chris Kurz, *Federal Reserve Bank*
Margaret Levenstein, *ISR, University of Michigan*
Jon Levin, *Department of Economics, Stanford University*
Nicole Mayerhauser, *BEA*
Kristen Monaco, *BLS*
Brian Moyer, *BEA*
Jeff Oldham, *Google, Inc.*
Trivellore Raghunathan, *SRC, University of Michigan*
Kamalesh Rao, *MasterCard*
Jerry Reiter, *Duke University*
Roberto Rigobon, *Sloan School of Management, MIT*
Scott Scheleur, *U.S. Census Bureau*
Matt Shapiro, *ISR, University of Michigan*
Chris Sparks, *BLS*
Erich Strassner, *BEA*
Steve Tae, *MasterCard*
Shaunda Villones, *BEA*

Simon Wilkie, *Microsoft, Inc.*

STAFF

Connie Citro, *Director, Committee on National Statistics*
Brian Harris-Kojetin, *Deputy Director, Committee on National Statistics*
Chris Mackie, *Senior Program Officer, Committee on National Statistics*
Anthony Mann, *Program Coordinator, Committee on National Statistics*
George Schoeffel, *Research Assistant, Committee on National Statistics*

Statistical Challenges in Combining Information from Big and Small Data Sources

Trivellore Raghunathan
University of Michigan

November 12, 2015

Abstract

Social Media, electronic health records, credit card transactional and administrative data, web scraping, and numerous other ways of collecting information have changed the landscape for those interested in addressing policy-relevant research questions. During the same time, the traditional sources of data, such as large-scale surveys, that have been a stable source for policy-relevant research have suffered setbacks due to large nonresponse and increasing cost of collecting such survey data. The non-survey data usually contain detailed information on certain behaviors on a large number of individuals (such as all credit card transactions) but very little background information on them (such as important covariates to address the policy-relevant question). On the other hand, the survey data contains detailed information on covariates but not so detailed information on the behaviors. Both data sources may not be perfect for the target population of interest. This paper develops and evaluates a framework for linking information from multiple imperfect data sources along with the Census data to draw statistical inference. An explicit modeling framework involving selection into the big data, sampling and nonresponse mechanism in the survey data, distribution of the key variables of interest and certain marginal distributions from the Census Data are used as building blocks to draw inference about the population quantity of interest.

1 Introduction

The digital revolution though at least 50 years old is coming to fruition now due to, in large part, ever increasing computational infrastructure and inexpensive storage. Social media, computerized or electronic records and many other digitized archives have changed the landscape of data. Statisticians have witnessed such changes in the landscape in the recent past. The advent of powerful desktop and server machines in the late eighties and early nineties made it possible to fit many statistical models that were impractical to implement a few years earlier and many old algorithms such as Metropolis (Metropolis et al (1953)), Hastings-Metropolis (Hastings (1970)), Gibbs sampling (Geman and Geman (1984)) and other Markov Chain Monte Carlo methods were no more theoretical exercises or relegated to main frame computers but became a common practice, so much so that, complex statistical model building has become quite routine.

The statisticians are at the cusp of the next stage of revolution where data from many sources and in many forms are becoming available and beckoning them to rise up to the challenge of integrating these data sources to construct inference about the population, their primary goal. The new challenge also includes the art and science of processing huge data sets that are not necessarily in the familiar rectangular format with rows for subjects and columns for variables.

During the same time period, the probability sample surveys, the traditional bread-and-butter tool for researchers has been facing challenges due to declining response rates. Many surveys conducted by survey research firms with tremendous perseverance and costs range between 40% to 50% and some telephone surveys much less. The government surveys are still eliciting larger response rates but at the enormous cost of nonresponse follow-up. Public fatigue, privacy and confidentiality concerns and costs will continue to affect the surveys. Hence, the surveys are relying more and more on post-survey adjustments using scant variables available on respondents and nonrespondents.

The real task for the statistical community is to face the challenge of declining response rates and the rising costs of conducting surveys with an increasing opportunity afforded by non-survey data sources without deviating from the principal objective: representative or "valid" inference about the target population of interest. There is a need for discovering a new set of tools or reshaping the old tools to leverage these two kinds of data sources. This can be done through refining the design of surveys and statistical models for combining information from multiple sources.

The goal of this paper is to lay out certain statistical framework for combining information from multiple data sources using the statistical modeling and imputation framework. Clearly lay out the assumptions needed to pool information from multiple sources and use those assumptions to construct "synthetic" or "plausible" data sets representative of the target population interest. This will enable the research community to broaden the scope of questions that can be asked and answered.

2 Big Data versus Survey Data

Declining response rates and increasing costs of traditional surveys and the advent of big data may tempt us to consider big data as the primary (or the only?) source for inferring about the population. To delve into the consequence of this possibility, consider the problem of estimating the prevalence rate, θ, of a certain attribute. Define a binary variable where $X = 1$ is for subjects with the attribute and $X = 0$, otherwise. A simple random sample survey of size n_S results in an estimate $\widehat{\theta}_S$, the sample proportion. The sampling variance of this estimate is $\theta(1-\theta)/n_S$. For now assume that there is no nonresponse.

Suppose that the same variable is captured in a non-survey data of size n_A, resulting in an estimate $\widehat{\theta}_A$, the proportion computed based the elements in the non-survey data. Suppose that $A = 1$ denotes that the person is captured in the non-survey data. Generally no information is available for the subjects not captured in the non-survey data. Nevertheless, let $Pr(A = 1|X = 0) = \pi$ and $Pr(A = 1|X = 1) = \rho\pi$ be the respective probabilities of capturing persons without and with the attribute. That is ρ is the rate of capturing a person with the attribute in the non-survey data relative to those without the attribute.

Suppose that we apply the same binomial model. Note that this a subjective model without the probability sampling framework as in the case of $\widehat{\theta}_S$. Some of the early references where such models were considered for non-probability samples are Smith (1983), Rubin (1987) and Deville (1991). The basic idea is to model the selection as a function of outcome and covariates and then lay out the conditions under which the observed sample can be used to project or predict the nonsampled part of the population. The response propensity models are examples of such subjective probability models that allows for post-survey adjustments (Little (1982)).

It follows that $Pr(X = 1|A = 1) = \theta\rho\pi/(\theta\rho\pi + (1-\theta)\pi) = \theta\rho/(\theta\rho + (1-\theta))$. The bias in the estimate $\widehat{\theta}_A$ is $-\theta(1-\rho)(1-\theta)/[1-(1-\rho)\theta]$. Thus the mean square error of $\widehat{\theta}_A$, under the assumed binomial model, is

$$MSE(\widehat{\theta}_A) = \frac{\theta(1-\theta)}{n_A}\frac{\rho + n_A\theta(1-\theta)(1-\rho)^2}{(1-(1-\rho)\theta)^2}$$

The relative efficiency of the estimate from the non-survey data relative to the random sample estimate is

$$RE_{A|S} = \frac{n_A(1-(1-\rho)\theta)^2}{n_S(\rho + n_A\theta(1-\theta)(1-\rho)^2)}.$$

Note that, this relative efficiency is not always greater than one even is n_A is very large compared to n_S. Let n_A be very large relative to n_S and the above equation simplifies to,

$$\frac{(1-(1-\rho)\theta)^2}{n_S\theta(1-\theta)(1-\rho)^2}.$$

An interesting question is when does the estimate from the big data become less efficient than the survey data. It can be shown the above equation is less than 1 for ($n_A >> n_S$), when

$$n_S \geq \frac{(1-(1-\rho)\theta)^2}{\theta(1-\theta)(1-\rho)^2}.$$

To get some perspective, suppose that $\rho = 1.2$ (that is, people with the attribute are 20% more likely to be captured in the non-survey data than those without the attribute) and the true prevalence rate is $\theta = 0.1$, then the non-survey data is less efficient whenever $n_S \geq 289$. Suppose that $\rho = 1.05$ and for the same θ, the threshold simple random size is $n_S \geq 4,489$. That is, the squared bias term tends to dominate even with the modest differential inclusion probabilities with respect to the outcome of interest in the non-survey data. It is not hard to imagine some differential inclusion probabilities related to the outcome of interest when the non-survey data are constructed for special purposes (Marketing companies, particular banks etc).

Of course, the real surveys rarely employ simple random sample design but typically involve unequal probabilities of selection, stratification and clustering. Thus, n_S could be interpreted as effective sample size adjusted for design effect.

The simple analysis suggests that selection bias can have a big impact on the inferences from the non-survey data and could not be even checked without having a reliable survey or some external data to check against or to calibrate. However, if an estimate, $\widehat{\rho}$, of ρ were available (say, based on a substudy) then one could construct a bias corrected estimate, $\tilde{\theta}_A$, by equating

$$\widehat{\theta}_A = \frac{\widehat{\rho}\tilde{\theta}_A}{1-(1-\widehat{\rho})\tilde{\theta}_A},$$

yielding,

$$\tilde{\theta}_A = \frac{\widehat{\theta}_A}{\widehat{\rho} + (1-\widehat{\rho})\widehat{\theta}_A}.$$

A pooled estimate combining the survey and non-survey data can be derived as

$$\widehat{\theta} = (v_S^{-1} + v_A^{-1})^{-1}(\widehat{\theta}_S/v_S + \tilde{\theta}_A/v_A)$$

where $v_S = \widehat{\theta}_S(1-\widehat{\theta}_S)/n_S$ and $v_A = \tilde{\theta}_A(1-\tilde{\theta}_A)/n_A$. An implicit Bayesian model is to treat $\theta|A \sim N(\tilde{\theta}_A, v_A)$ as the prior distribution and $\widehat{\theta}_S|\theta \sim N(\theta, v_S)$ as the sampling distribution.

3 Strategies for Estimating Selection Bias

It is critically important to assess and estimate the selection bias term ρ. Fortunately, the modeling framework provides for laying out the assumptions and some approaches for estimating the selection bias. Suppose that Z is a covariate with k categories such that $Pr(A = 1|X = 1, Z = j) = Pr(A = 1|X = 0, Z = j) = Pr(A = 1|Z = j)$, $j = 1, 2, \ldots, k$. This is akin to missing at random assumption in the missing data framework (Rubin (1976)) conditional on Z.

Note that

$$Pr(A = 1|X = 1) = \sum_j Pr(A = 1|Z = j)Pr(Z = j|X = 1)$$

and

$$pr(A = 1|X = 0) = \sum_j Pr(A = 1|Z = j)Pr(Z = j|X = 0)$$

. Writing

$$Pr(A = 1|Z = j) = Pr(Z = j|A = 1)Pr(A = 1)/Pr(Z = j),$$

we obtain

$$\rho = \frac{Pr(A = 1|X = 1)}{Pr(A = 1|X = 0)} = \frac{\sum_j Pr(Z = j|X = 1)Pr(Z = j|A = 1)/Pr(Z = j)}{\sum_j Pr(Z = j|X = 0)Pr(Z = j|A = 1)/Pr(Z = j)}$$

Thus, to implement this method we need estimates of the marginal and various conditional distributions of the covariate, Z.

- From the non-survey data we need estimates of $Pr(Z = j|A = 1)$.

- The Census or the population data may provide $Pr(Z = j)$
- A sample survey or a pilot study may provide $Pr(Z = j|X = l), l = 0, 1$.

The categorical nature of the covariates makes these building blocks as aggregate data that producers of non-survey data may be able to provide without violating privacy and confidentiality. For example, if the non-survey data source is a bank, for example, and Z is the categories of total "volume", then the bank may be able to provide the marginal distribution of based on its customers.

What are some of the options for constructing Z? Suppose that the non-survey and survey data have some common covariates U. Suppose that $\widehat{\beta}_S$ is the estimated regression coefficient, in a logistic regression model with X as the dependent variable and U as independent variables, obtained from the survey data. Let $Z = [1 + \exp(-U^t\widehat{\beta}_S)]^{-1}$ be the predicted probability. The same regression coefficient, $\widehat{\beta}_S$, is then applied to the non-survey data to construct Z. That is, Z is the (counterfactual) prediction of X for the subjects in the non-survey data that would have been obtained had they been in the survey data. The underlying assumption is that conditional on having the same prediction under the survey data, the actual attribute status is not related to the selection into non-survey data. The predicted variable, Z, can be categorized to create classes.

The second approach is to use some common variables between the non-survey data and the sample frame data. Some examples of such variables are block or block group characteristics. Suppose that $S = 1$ indicates a sampled subject and $S = 0$ indicates a non-sampled subject. Let U be the frame variables also available in the non-survey data (or can be attached to non-survey data). Let $\widehat{\beta}_S$ denote the regression coefficient from the logistic regression model predicting S from U. Apply this estimated regression coefficient to the non-survey data. This covariate represents the likelihood of subjects in the non-survey data for being predicted to be in the sample. Again, this covariate could be categorized to form classes.

A final example of a strategy for constructing the covariate Z is the propensity of being in the survey data. Specifically, append the non-survey and survey data and define $D = 1$ for the survey subjects and $D = 0$ for the non-survey subjects. Estimate the propensity score by using a logistic regression model with D as the dependent variable and all the common covariates in the two data sets. The categories can be created based on the propensity score. The rationale underlying this strategy is that if the subject in the non-survey data matches to subject in the survey data then the labeling of subjects as survey/non-survey is completely at random.This strategy was used to correct for discrepancies between the self-report and clinical measures of chronic conditions such as hypertension, diabetes etc by pooling data from the National Health Interview Survey and National Health and Nutrition Examination Survey as described in Schenker, Raghunathan and Bondarenko (2010).

The central theme of all these approaches is to balance or match the non-survey data with the survey or population data through propensity scoring. Within the matched sets, selection bias is assumed to be non-existent or at least negligible. Note that, the bias corrected non-survey data estimate will have very small mean square error relative to survey based estimate (if the bias correction is successful). Thus, the the survey goal could be just to provide enough data to permit bias-correction. A smaller scale survey with high response rate could possibly be mounted with lower cost and thus leveraging the information in the larger non-survey data.

Obviously, the survey data is subject to nonresponse but several studies (Groves et al (2010)) have shown that even with high nonresponse rate, the survey estimates suffer from lower nonresponse bias. Furthermore, auxiliary variables can be collected on respondents and nonrespondents (through proper planning at the design stage and collected during the conduct of the survey), post-stratification techniques can be used derive unbiased estimates from the survey data.

Some sources of auxiliary variables include interviewer observations, contextual or geographical data estimated from a variety of sources, commercial data etc. To some extent, the survey world did not creatively plan the collection of auxiliary variables with an anticipation of the steep decline in the response rates. One of the reason is that survey inference world was less embracing towards the use of statistical modeling in the inferential activities where as the non-survey inference world fully embraced and exploited the modern statistical modeling and computational advances to its great advantage. The quote "All models are wrong and some are useful", attributed to George Box, a famous statistician summarizes the attitude needed: Carefully craft the model that captures the important features of the data being analyzed, perform proper diagnostics to assess the model fit and then proceed with the inference about the population, fully incorporating the uncertainties in the non-observed data conditional on the model and an assessment of sensitivity of the inferences to the model assumptions.

4 Going beyond each through combining

Consider a situation were a data source A provides variables (U, X, Y), the data source B provides (U, X, Z) and data source C provides (U, Y, Z). If the data sources A, B and C are representative of the same population then vertically appending the data creates a traditional missing data problem (missing Z in the data set A, missing Y in the data set B and missing X in the data set C). Existing technology such as multiple imputation can be applied to create completed data sets that allows joint analysis of (U, X, Y, Z). Note that, such leveraging extends the utility of each data source beyond what it was intended to be. This strategy could be used by the data repositories, Federal agencies to use the variety of data already collected, harmonize the variables and link it spatially and temporally.

An example of one such project is to consider the 1940 census which is now available electronically to create a cohort of individuals and then try to link (deterministic or probabilistic) first to all available digitized information such as the Current Population Surveys, American Community Surveys, various other surveys, Administrative records, mortality files etc. This requires a concerted efforts working across agencies within the confines of secured environment, such as Census Bureau Research Data Center. This first stage effort will provide a data set with considerable holes (missing information). The investigation of missing portions will then lead to sampling of non-digitized records such as later year census data for digitization and incorporation into the data set.

Obviously, the cohort formed from the 1940 census is not a representative for the later years. Thus, sub-sampling and digitization of subjects in the later census years and and attaching available survey data to them will improve the representativeness and provide better temporal picture. Once all reasonable efforts have been made to fill-in as much information as possible through deterministic or probabilistic linking then one can adopt a statistical approach for multiply imputing the missing portions of the data set. Thus creating a retrospective observation based longitudinal data entirely by leveraging the existing data resources.

The goal is not to create an actual data set, but a plausible data set that matches the population in various respects. Just like an imputed data for any one survey is not an actual data set but a plausible data set. The reasonableness of such a data set can be assessed by comparing the inferences from this data set to the inferences from the actual data set for a given time period and given set of variables. For example, one can check whether the plausible data set so constructed yield descriptive and analytical inferences for, say the year 1990, yield similar to the one based on, say 1990 long form. Such calibration of the plausible data increases the confidence in the inferences constructed from it.

Returning the example with three data sources, A, B and C, suppose that each one them may be subject to selection bias. Usually, the selection bias is not be known. The unknown information are the conditional distributions, $[Z|U, X, Y, A]$, $[Y|U, X, Z, B]$ and $[X|U, Y, Z, C]$.

Suppose that a small representative survey is conducted to collect data D, on (U, X, Y, Z), appropriately weighted and imputed for missing values using the design variables, paradata and other auxiliary variables. The goal is not make this survey a primary vehicle for drawing inference about the population but enough to estimate the quantities needed to leverage the large data sets A, B and C.

The following strategy could be used to achieve our goal of creating a plausible data set from the population:

1. Append all four data sets (vertically concatenate). Create a categorical variable V with three levels, with $V = 1$ for data A, $V = 2$ for data B and $V = 3$ for data C. Set V to be missing for all subjects in the data set D. When this variable is imputed the observed data is being used allocate subjects in the data set D to one of the three data sources.

2. Impute the missing values in Z for data A by applying the restriction that model be fit and predicted values be generated by sub setting the data with $V = 1$.

3. Impute the missing values in Y in the data B by restricting model fit and imputation to $V = 2$

4. Impute the missing values in Z in the data set C by restricting the model fit and imputation to $V = 3$.

5. The final step is assign weights to subjects in the data sets A, B and C commensurate with their representation in the population. For example, the post-stratification based on the population characteristics (for example, the census data or estimated from large surveys such as the American Community Survey, the Current Population Survey or the National Health Interview Survey). The second option is to use the imputed data set D to estimate the representation of the subjects like those in A, B and C. Suppose that p_A, p_B and p_C be the weighted estimated of proportion for categorical variable V in the data set D. Let m_A, m_B and m_C be the sizes of data sets A, B and C, respectively

with $m = m_A + m_B + m_C$. Assign each subject in the data set A the weight of m_A/mp_A. Similarly, m_B/mp_B and m_C/mp_C for the data sets B and C, respectively. All subjects in the data set D receives the original survey weight.

To incorporate the uncertainty in the imputations, the above steps can be repeated several times to create a set of multiply imputed plausible data sets. Standard multiple imputation combining rules (Rubin (1987), Little and Rubin (2002), Raghunathan (2015)) can be applied to create inferences. Of course, this strategy extends to many variables with arbitrary pattern of missing data and more than three data sources that can be pooled to create large plausible data set from the population adjusted for selection bias.

One of the ongoing project involves creating an infrastructure to develop an understanding of relationship between demographic and socio-economic factors (X), health conditions (D) and medical expenditures (E). Each of these variables are multivariate. Unfortunately, there is no single data source that provides comprehensive information on all three domains for the entire population. However, there are several data sets measuring a subset of these domains. For example, the Medicare Current Beneficiary Survey, National Health Interview Survey, National Health and Nutrition Examination Survey, Health and Retirement Study, Medical Expenditure Panel Survey, National Comorbidity Survey etc are some of the representative surveys provide data in some of these dimensions. Through calibration, post-stratification and imputation plausible data set is being created for four age segments of the population: Age 65 and above, 45 to 64, 18 to 44 and under 18 years of age. The work has been completed for Age 65 and above for the period 1999-2009, primarily using MCBS, NHANES and CMS claims as data sources (Cutler et al (2015)).

5 Discussion

Combining survey and non-survey data sources provides unique opportunities to extend the usefulness of each data source and pose challenges in terms of the methodology to be used to harness information from these sources. The declining response rates in sample surveys and potential selection bias in the non-survey data sources makes the task as that of pooling information from imperfect sources.

Even with low response rate surveys, through auxiliary variables and post-stratification, it is possible to adjust for bias and by reducing the sample size, more efforts can be devoted for increasing the response rate or reducing the nonresponse bias. This smaller high quality survey can then be used to correct for potential selection bias in the non-survey data.

The central theme of this paper is that task of combining information from multiple imperfect data sources can be accomplished through proper development of statistical models with reasonable assumptions that be directly or indirectly tested or validated. The current missing data framework, modeling and software can be modified to achieve this goal. Some simple examples given in this paper are just for kindling the imagination for this line of research to be undertaken by the scientific community.

There are several limitations. The data sources could be collected under different contexts, some are self-reports and others could be record based. It is possible that some data were collected on web, some on telephone, some through mail and some through in-person interview. The mode differences may make the measurement not comparable. There may design differences across the surveys being pooled. All these are challenges that require thoughtful small scale experiments incorporation of those information through modeling. Finally, the landscape for the data analysis has changed and this reality should force us to think creative ways to harness the information from survey and non-survey data sources.

References

[1] Deville, Jean-Claude (1991), A theory of quota surveys, *Survey methodology*, 17, 163-181.

[2] Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayeisan restoration of images, *IEEE Transactions on Patern Analysis and Machine Intelligence*, 6, 721-741.

[3] Groves, R. M. (2008). Nonresponse rates and nonresponse bias in household surveys, *Public Opinion Quarterly*, 70, 646-675.

[4] Hastings, W. K. (1970). Monte carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97-109.

[5] Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T.M., Casas-Cordero, C, Lemay, M., Peytchev, A., Groves, R.M., and Raghunathan, T.L. (2010). Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Nonresponse: Examples from Multiple Surveys. *Journal of Royal Statistical Society (Series A)*, 173, 389-407.

[6] Little, R. J. A. (1982). Models for Nonresponse in Surveys, *Journal of American Statistical Association*, 77, 237-250.

[7] Little, R. J. A. and Rubin, D. B. (2002). **Statistical Analysis with Missing Data**, New York: Wiley.

[8] Metropolis, N. Rosenbkuth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast compouting machines. *Journal of Chemical Physics*, 21, 1087-1092.

[9] Raghunathan, T. E. (2015). **Missing Data Analysis in Practice**. CRC Press: Boca Raton, Florida.

[10] Rubin, D. B. (1987). **Multiple Imputation for Nonresponse in Surveys**, Wiley: New York.

[11] Schenker, N. and Raghunathan, T. E. (2007). Combining information from multiple surveys to enhance estimation of measures of health, *Statistics in Medicine*, 26, 1802-1811.

[12] Schenker, N., Raghunathan, T. E., and Bondarenko, I. (2010). Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey, *Statistics in Medicine*, 29, 533-545.

[13] Smith T. F. M. (1983). On the validity of inferences from Non-random samples, *Journal of the Royal Statistical Society*, 146, 394-403.

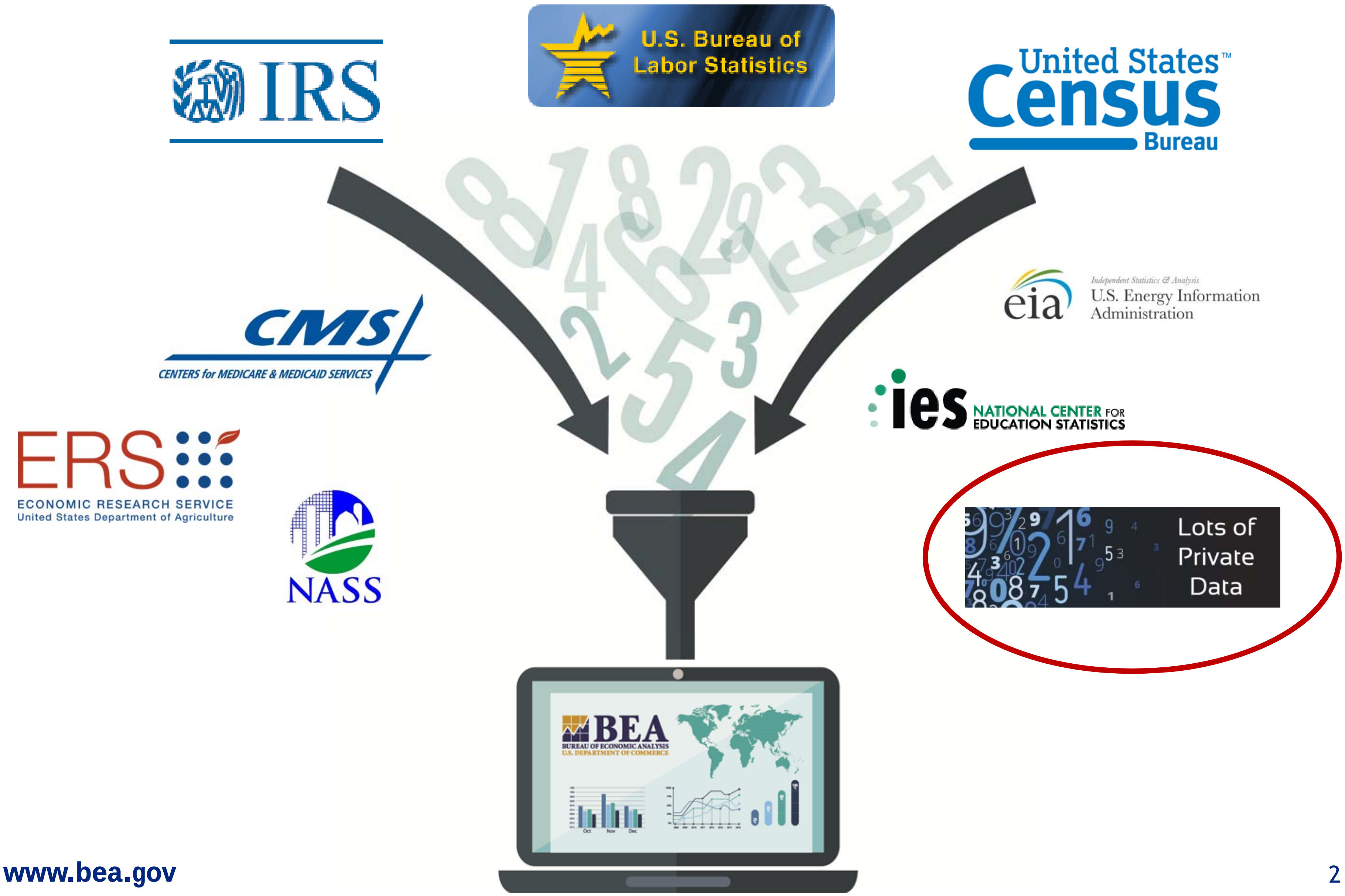
Commercial Big Data and Official Statistics

David Johnson

Federal Economic Statistics Advisory Committee

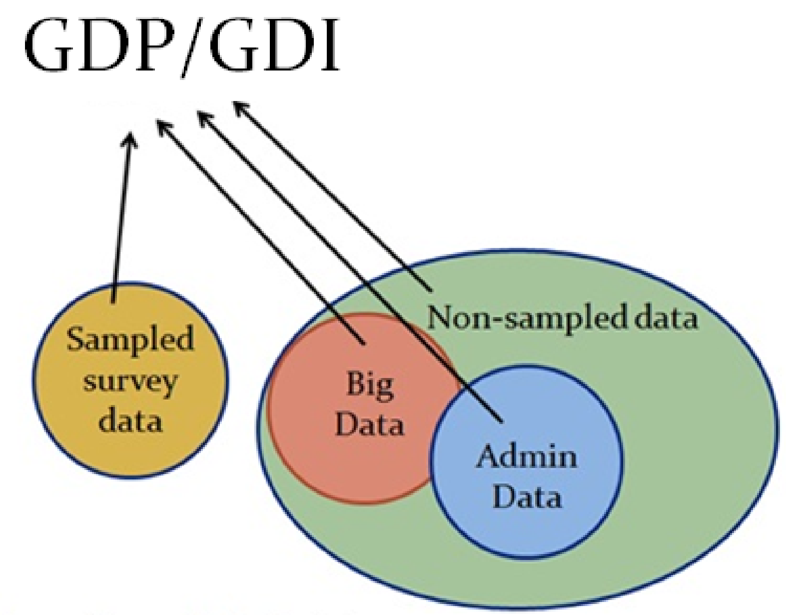
June 12, 2015

BEA Uses a Variety of Data



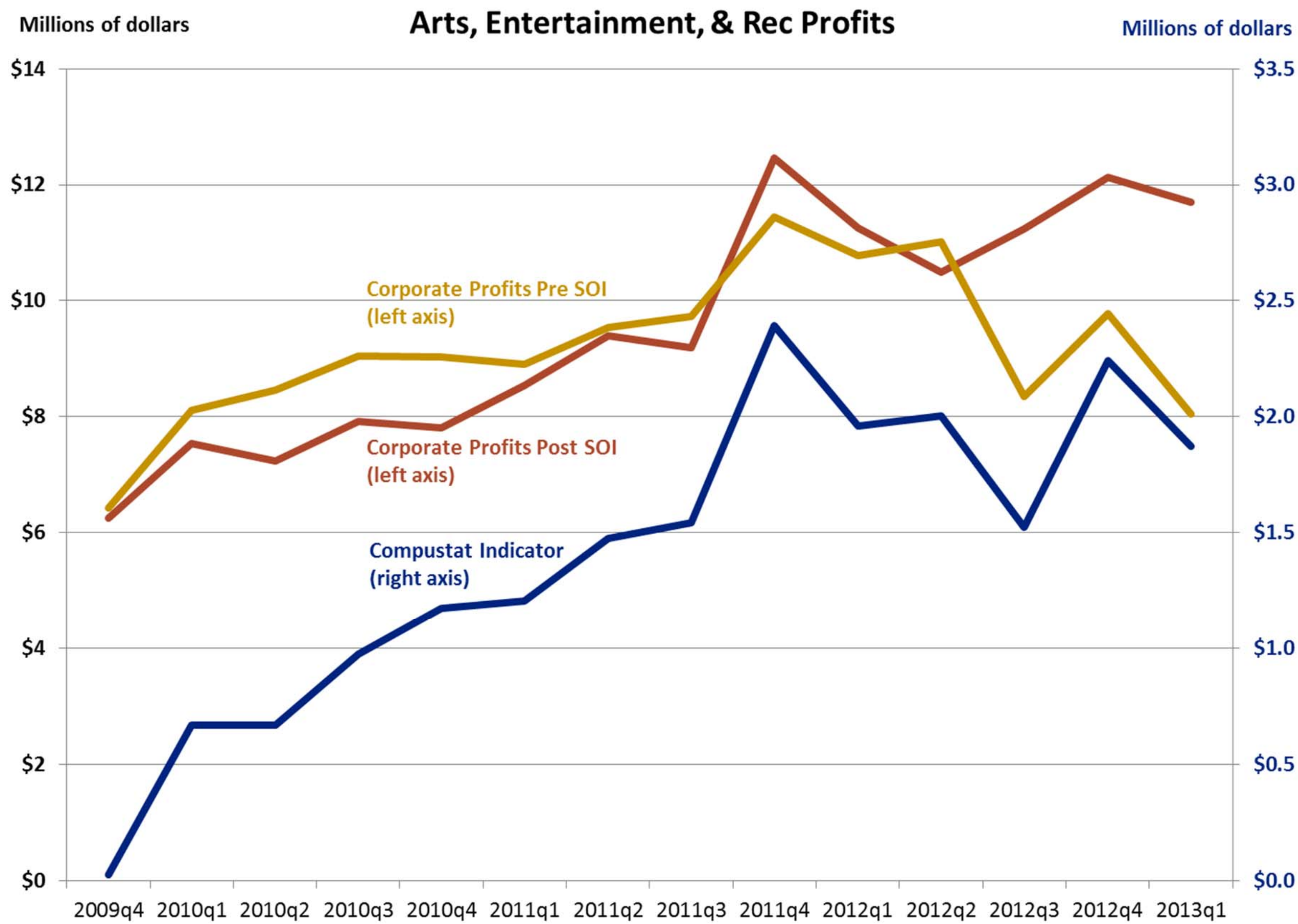
Current Uses of Commercial and Administrative Data

- BEA has a history of using nontraditional data for estimation (over 120 data sources)
- Current private source data include:
  - Ward's/JD Powers/Polk (auto sales/price/registrations)
  - Compustat (profits)
  - American Petroleum Institute (oil drilling)
- Current administrative source data include:
  - IRS, Statistics of Income
  - DOL, Unemployment Insurance data
  - FDIC, Commercial bank assets and liabilities data



Source: Bureau of Labor Statistics

Example: using Compustat for Corporate Profits



Challenges Using Commercial Data

- How representative are the data?
- Do the concepts match those needed for national accounts?
- Do the data provide consistent time series and classifications?
- Is it possible to fill gaps in coverage?
- How timely are the data?
- How cost effective?

BUT…

Opportunities for Commercial Data

- Provide indicators and extrapolators
- Fill data gaps
- Expand geographic detail
- Confirm trends
- Aid in future research efforts
  - E.g., distributional measures

Questions for Discussion

- How could the new data be used in official economic statistics?
  - Mint Bills
  - Paycycle and QuickBooks
  - CFPB Consumer Credit Panel
- Have the estimates from these new data been compared to official estimates?
- Are there suggestions for other possible data sources?
- What are the challenges in allowing agencies to access and use these data?

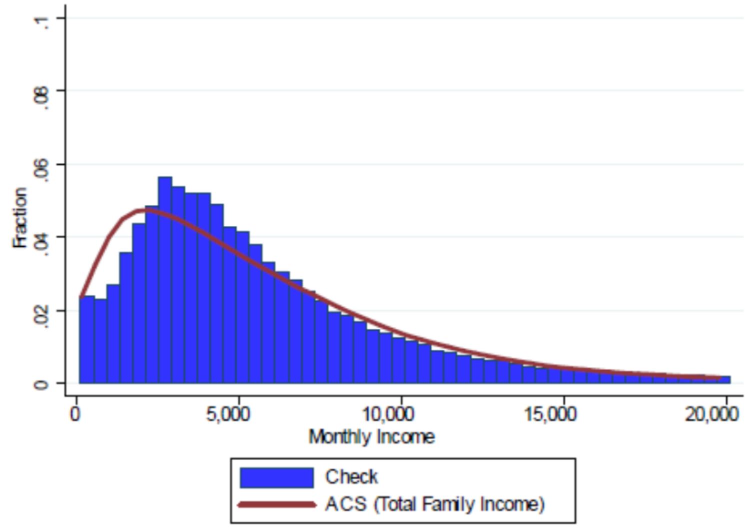
Mint Bills data are not representative of population – how does this affect measures

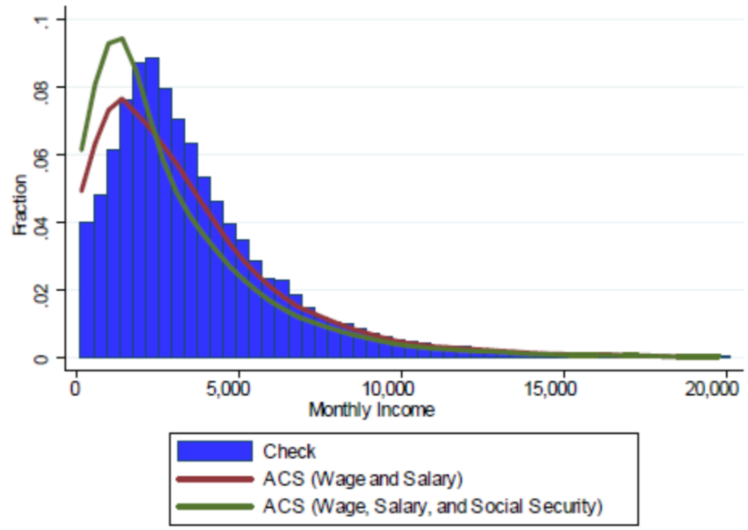
Table 1: Mint Bills vs. ACS Demographics

	Mint Bills	ACS
Female	40.07	51.41
Age		
18-20	0.59	5.72
21-24	5.26	7.36
25-34	37.85	17.48
35-44	30.06	17.03
45-54	15	18.39
55-64	7.76	16.06
65+	3.48	17.95
Highest degree		
Less than College	69.95	62.86
College	24.07	26.22
Graduate School	5.98	10.92
Census Bureau Region		
Northeast	20.61	17.77
Midwest	14.62	21.45
South	36.66	37.36
West	28.11	23.43

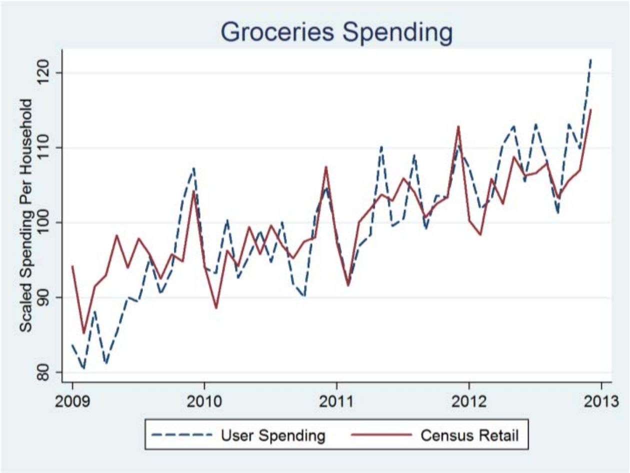
Note: The sample size for Mint Bills is 59,072, 35,417, 28,057, and 63,745 for gender, age, education and region, respectively. The sample size for ACS is 2,441,532 for gender, age, and region, and 2,158,014 for education

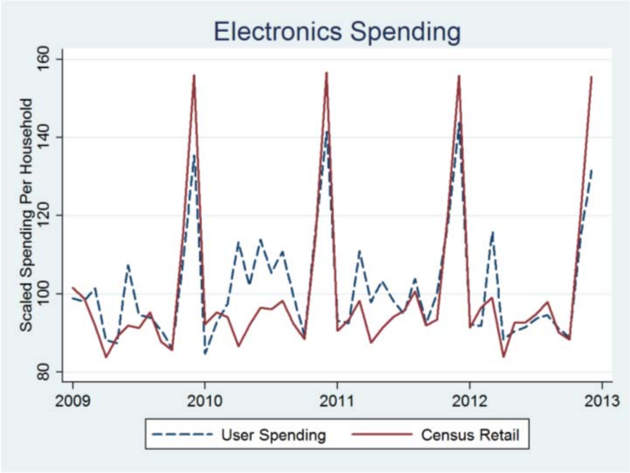
And income distribution is less skewed

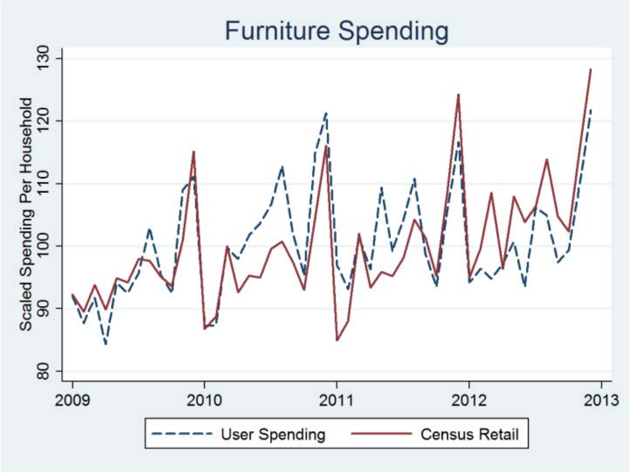




Need detailed spending to compare Mint Bills data to Census retail data, as in Baker (2014)

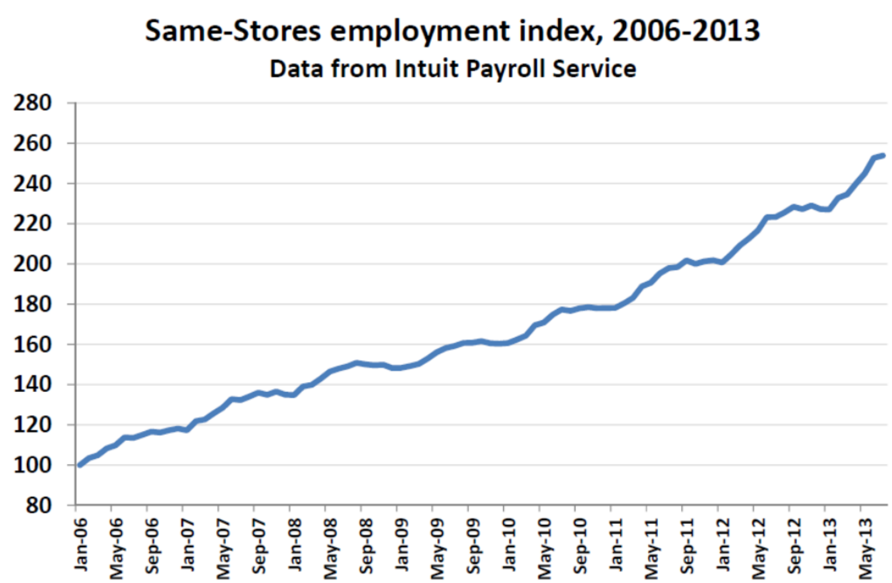


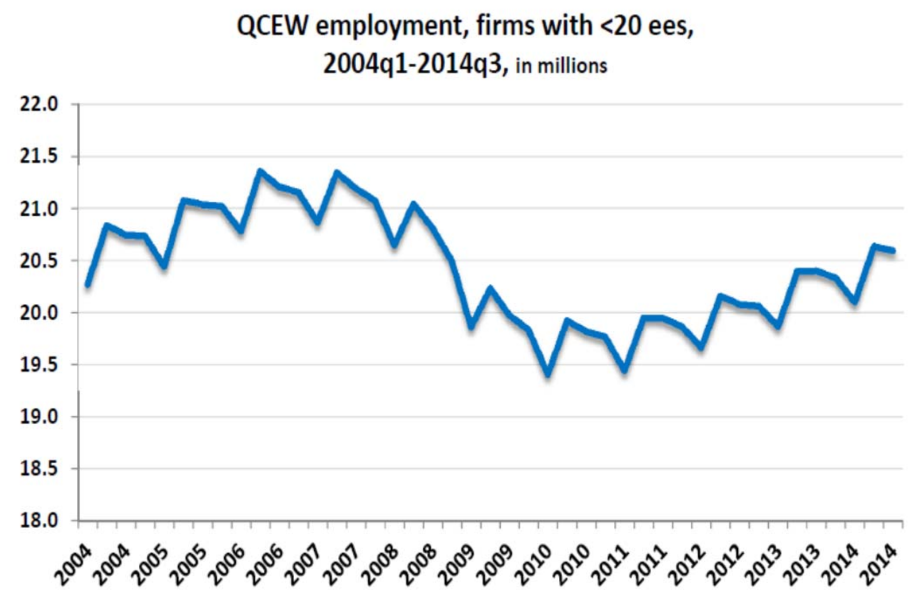




Source: "Debt and the Consumption Response to Household Income Shocks," S. Baker, Kellogg School of Management, 2014

Why do the Paycycle small business growth rates differ from QCEW?





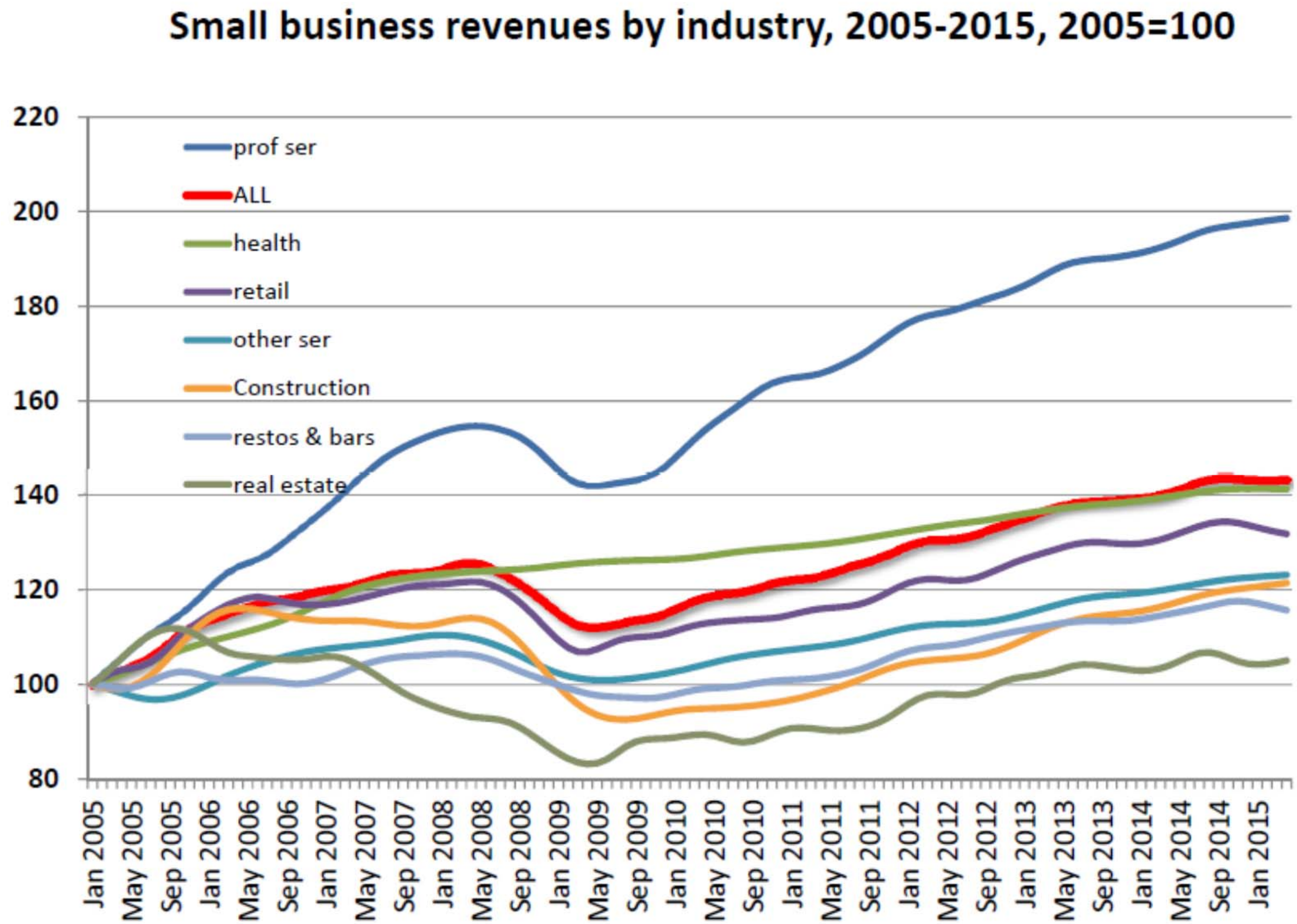
BEA initiative
Big Data for Small Business

Encourage small business growth

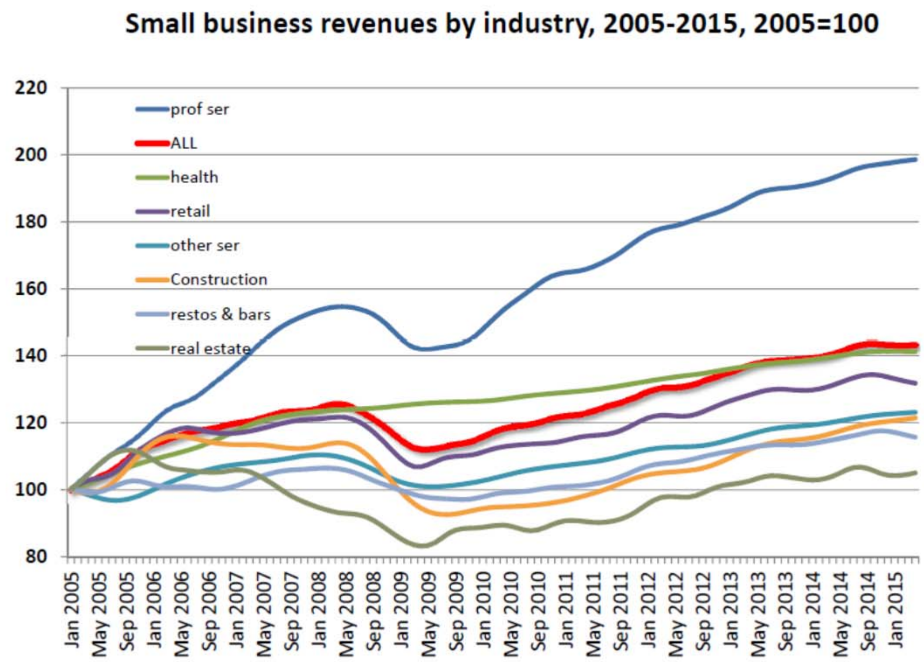
- Expanded information on small businesses would support the Department's and the Administration's goal to grow this important sector

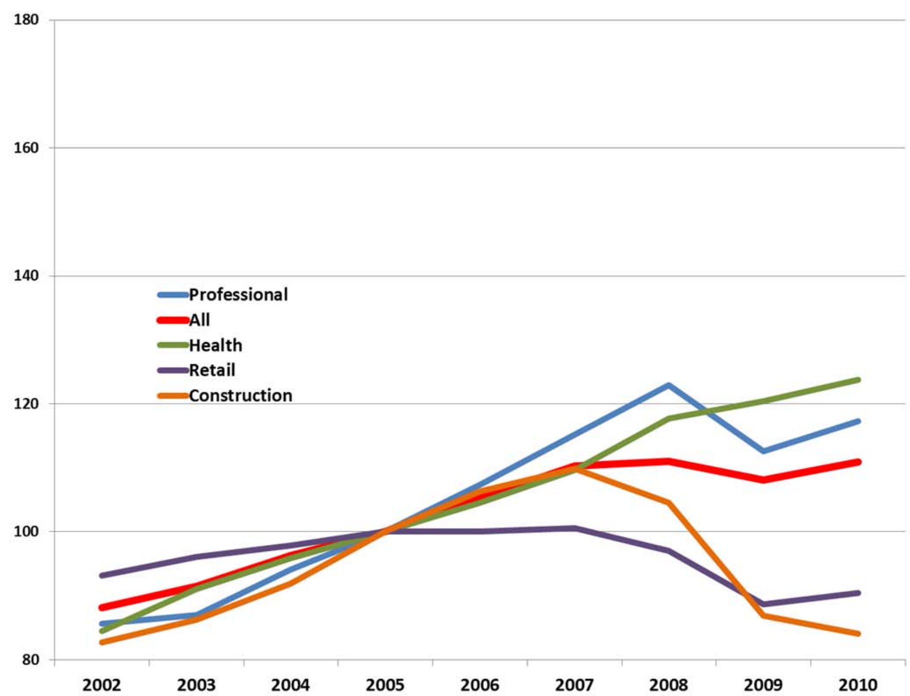
- Initiative will develop a new small business satellite account that would comprise:
  - Small business GDP
  - Small business GDP broken out by industry and by regions of the country
  - Distributional information of the employment and sales of small businesses
  - Information on the legal form, taxes, and net income of small businesses

Quickbooks data may provide insight into small business revenues



Need to Compare estimates to other sources, e.g. Small Business Administration





Source: "Small Business GDP: Update 2002-2010," Kathryn Kobe, Small Business Administration, 2012

How representative is the CFPB's Consumer Credit Panel (CCP)…

…a longitudinal sample of approximately 5 million de-identified credit records that is nationally representative of the credit records maintained by one of the nationwide credit reporting agencies (NCRA).

TABLE 1: EFFECTS OF SAMPLE EXCLUSIONS AND INCLUSIONS

	Scored Records	Credit Invisibles	Stale-Unscored	Insufficient-Unscored
Baseline Estimate	188.6	26	9.6	9.9
Exclusions:				
Missing in 2014 (Total)	-4.3	+11.7	-1.7	-5.7
Observed Merge	-3.0	+6.6	-1.2	-2.4
Disappeared	-1.2	+5.0	-0.5	-3.3
Missing Age	-6.5	+7.4	-0.4	-0.5
Inclusions:				
Debt Collection Only	+0.1	-2.1	+0.09	+1.9
Public Record Only	+0.01	-0.4	+0.01	+0.3

Source: "Data Point – Credit Invisibles," CFPB Office of Research, 2015

Need to compare the characteristics to another data source, as does the NY Fed CCP

Table 2. Comparison of 2008 Age Distributions Based on Extended Sample

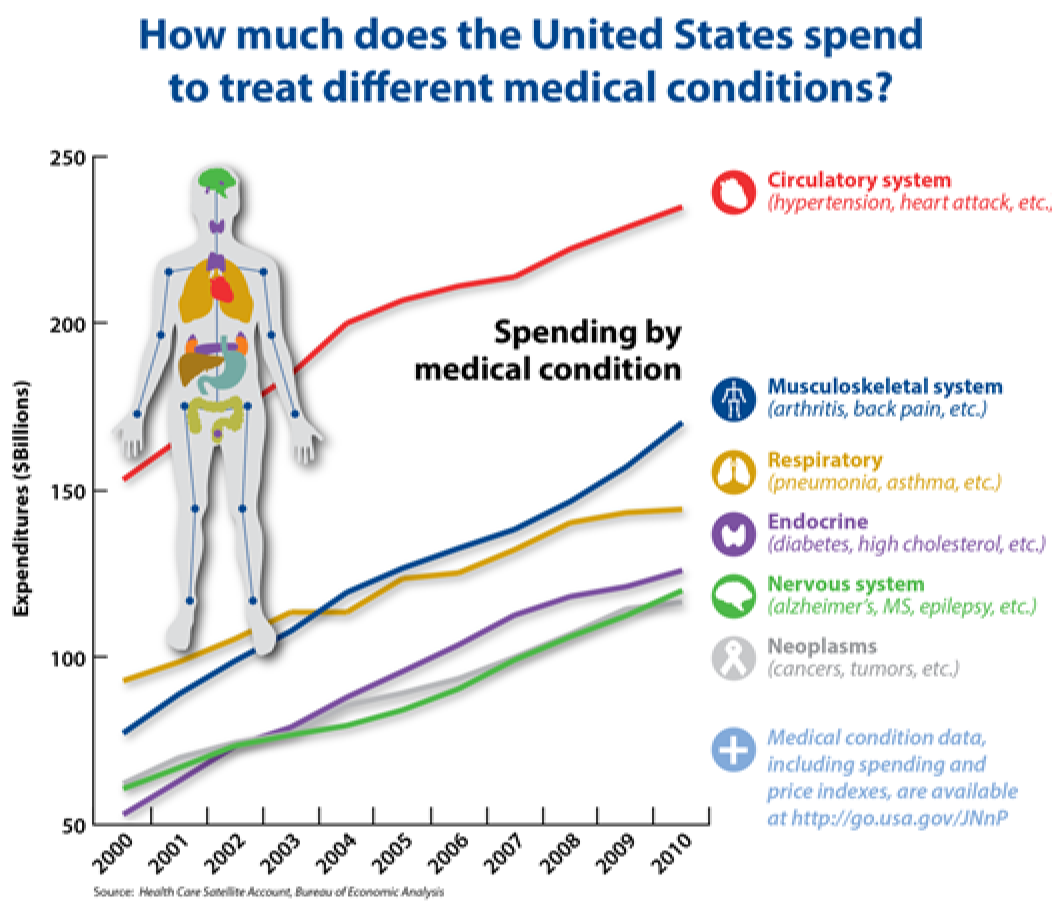
Age	US			NY State			NYC			Manhattan		
	ACS Age≥18	ACS ≥20	FRBNY All	ACS Age≥18	ACS ≥20	FRBNY All	ACS Age≥18	ACS ≥20	FRBNY All	ACS Age≥18	ACS ≥20	FRBNY All
18-24	13.1	9.5	9.4	13.3	9.7	9.1	12.6	9.3	8.9	10.5	7.9	7.4
25-34	17.6	18.3	17.0	16.8	17.5	17.1	19.2	19.9	21.6	21.6	22.[illegible]	24.4
35-44	18.5	19.3	18.4	18.6	19.4	18.9	20.5	21.2	20.8	23.4	24.0	21.2
45-54	19.3	20.1	19.6	19.4	20.2	19.9	18.2	18.8	19.2	16.6	17.1	17.0
55-64	14.7	15.3	15.5	14.7	15.4	15.3	13.6	14.1	14.2	12.7	13.0	13.9
65-74	8.7	9.1	9.5	8.8	9.2	9.1	8.3	8.6	7.8	8.1	8.3	8.0
75-84	5.7	5.9	6.7	5.8	6.1	6.6	5.4	5.6	4.8	5.2	5.3	4.9
85+	2.4	2.5	4.0	2.6	2.7	4.0	2.3	2.4	2.8	2.3	2.4	3.2
Total (millions)	230.2	221.1	266.2	15.1	14.5	16.6	6.4	6.2	6.8	1.4	1.3	1.7

American Community Survey figures are 1-yr estimates for 2008 from tables B11016 Household Type by Household Size. The FRBNY figures are based on the Q4 2008 wave in the consumer credit panel. All counts are in millions.

Source: "An Introduction to the FRBNY Consumer Credit Panel," D. Lee and W. van der Klaauw, Federal Reserve Bank of NY Staff Report, 2010

Example: Health Care Satellite Account

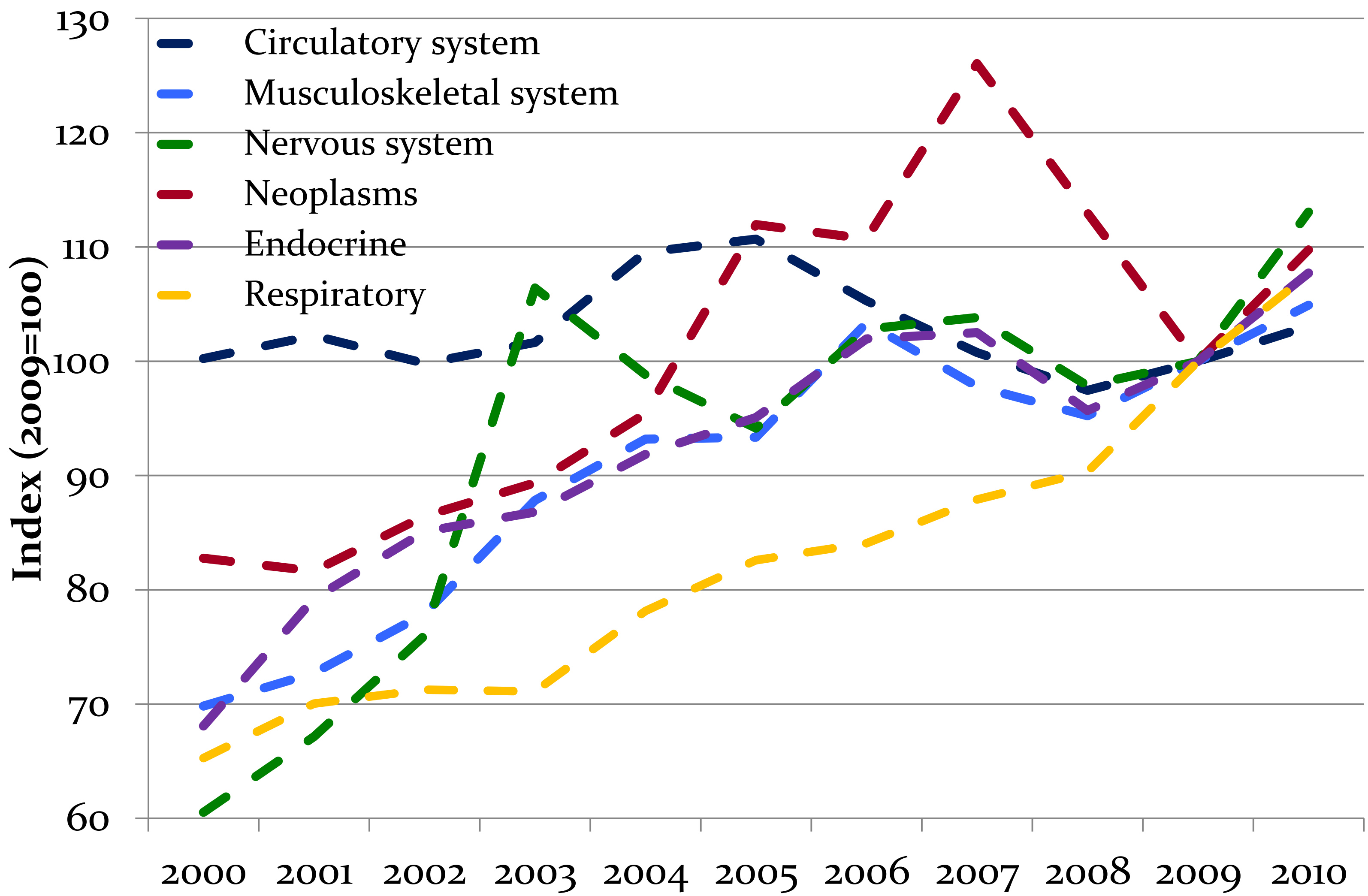
- Annual statistics for 2000-2010 that provide information on spending and price changes by disease category
- BEA combined billions of claims from both Medicare and private commercial insurance to determine the spending for over 250 diseases



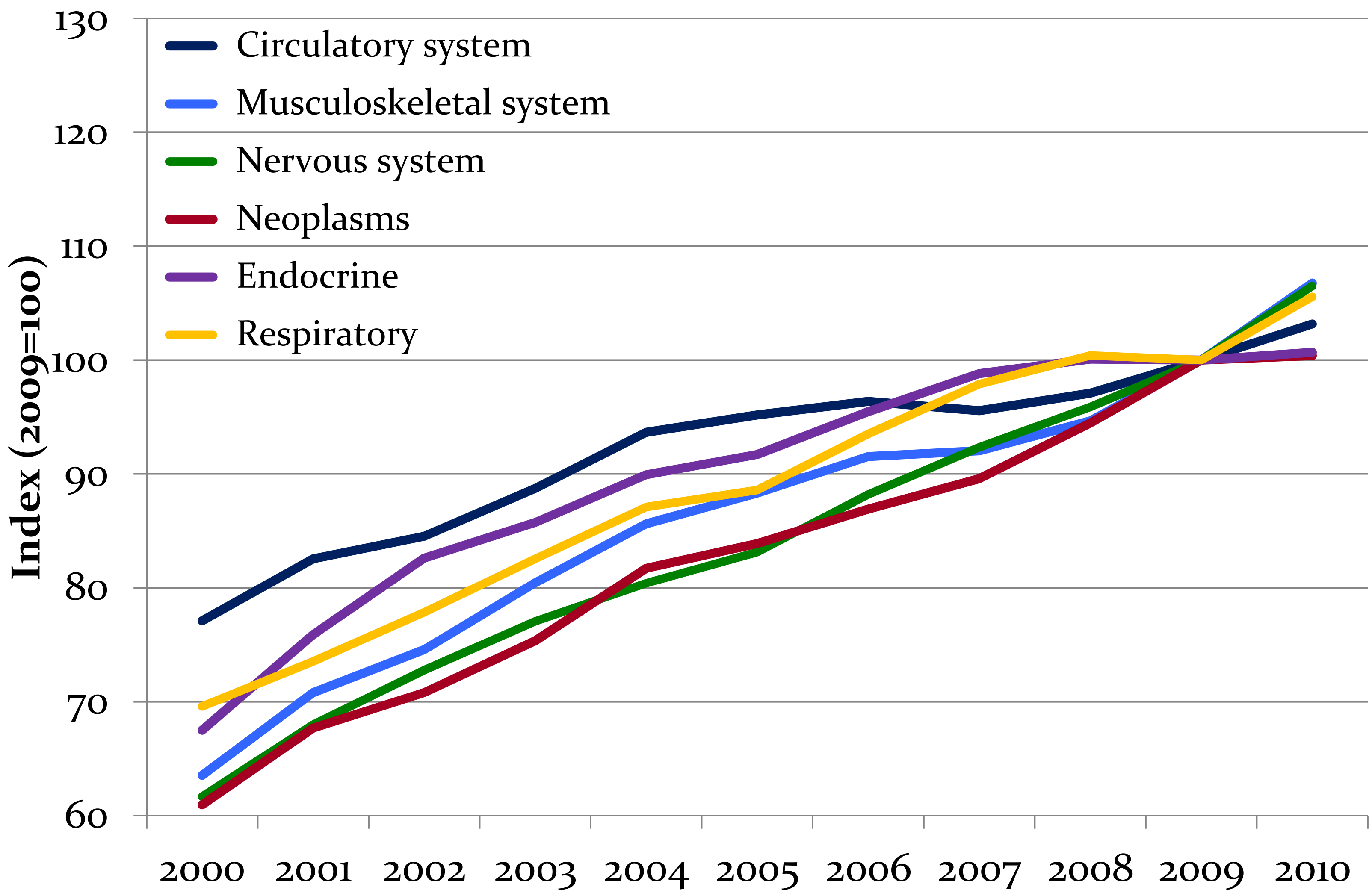
Construction of Blended Account

- Use survey population weights to fold in data from different sources

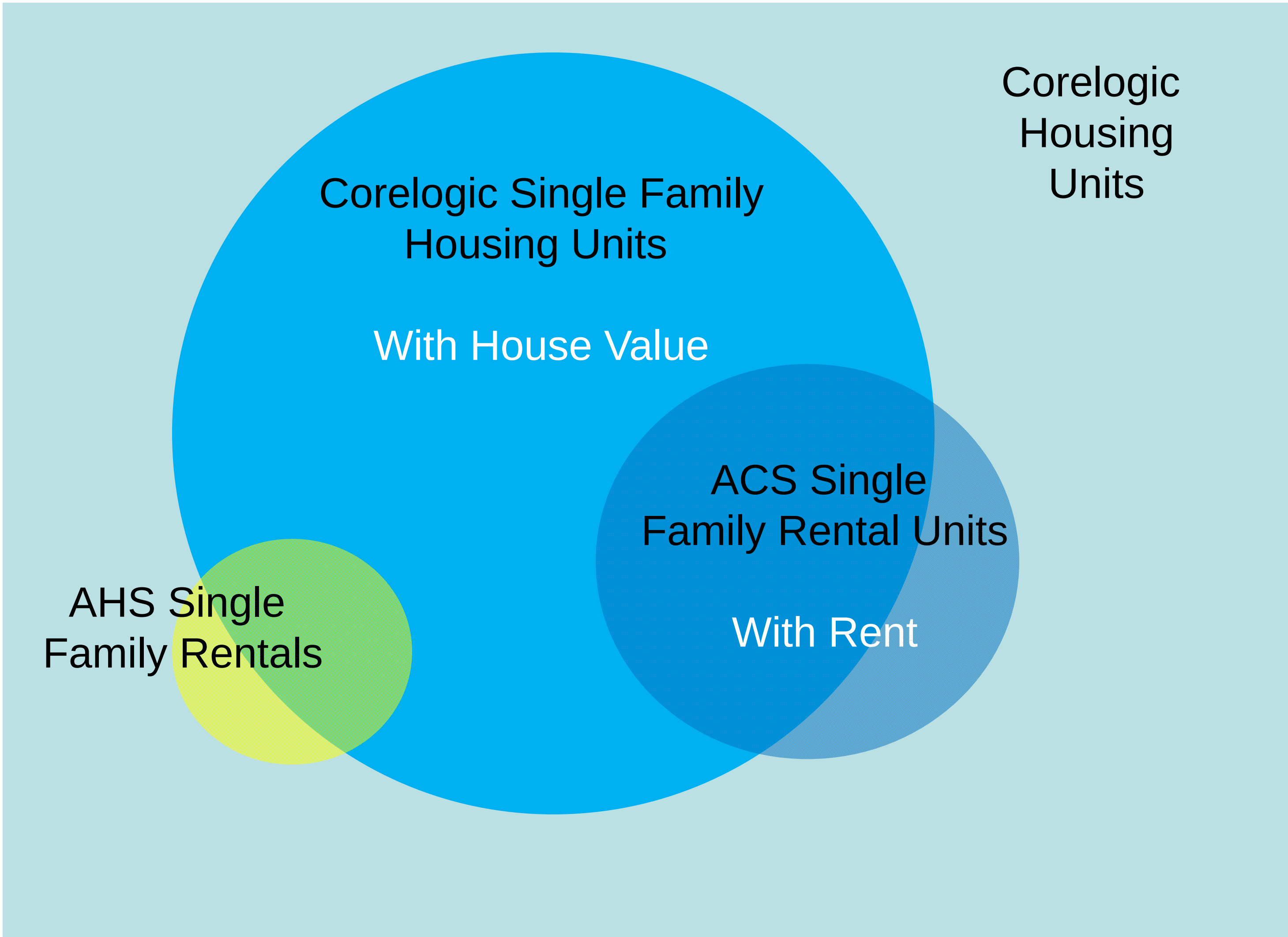
Health Care Satellite Account: Survey Data Only



Health Care Satellite Account: Survey + Big Data



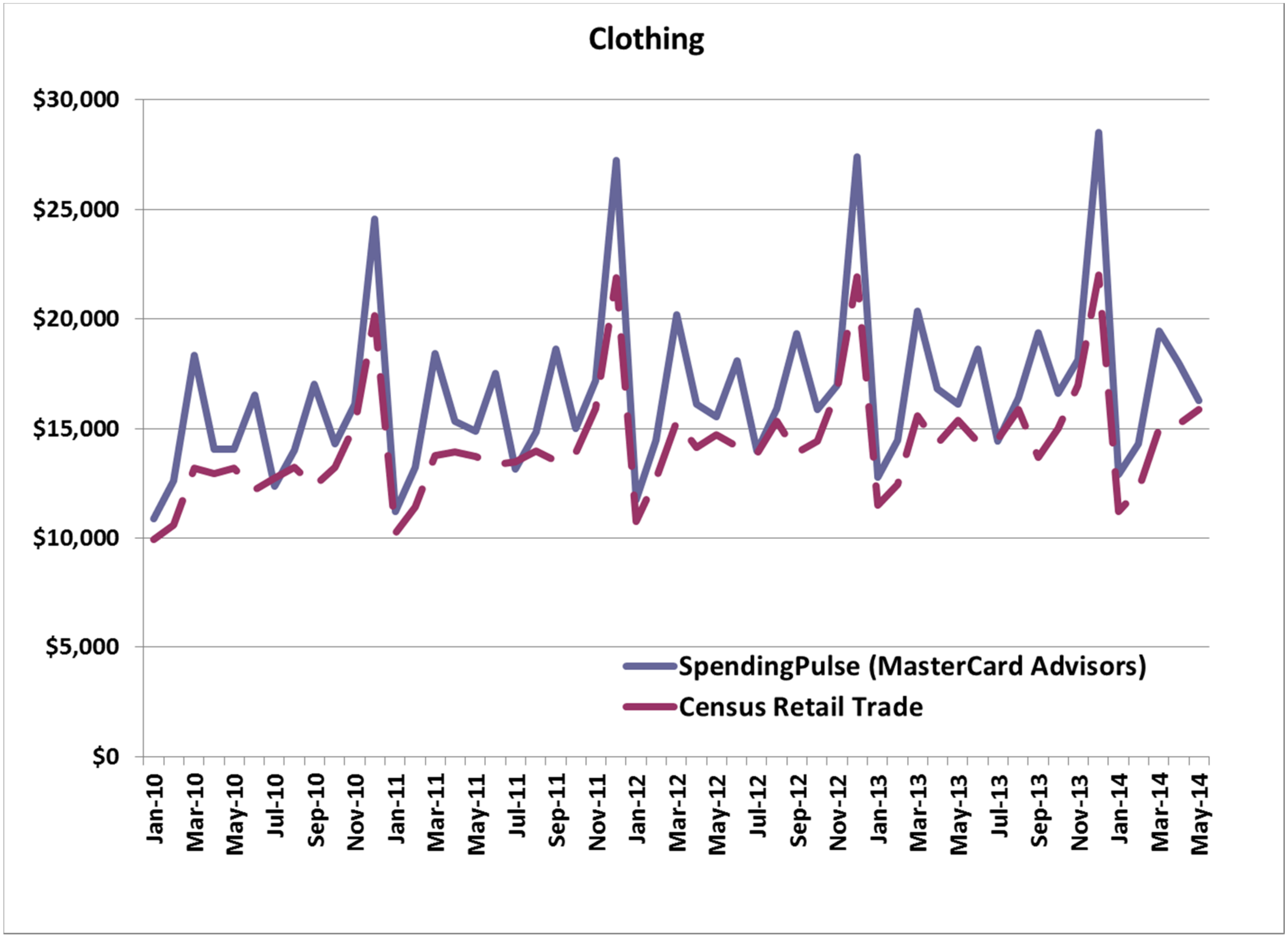
Example: Match Corelogic house prices to ACS rents to obtain rent to value ratio



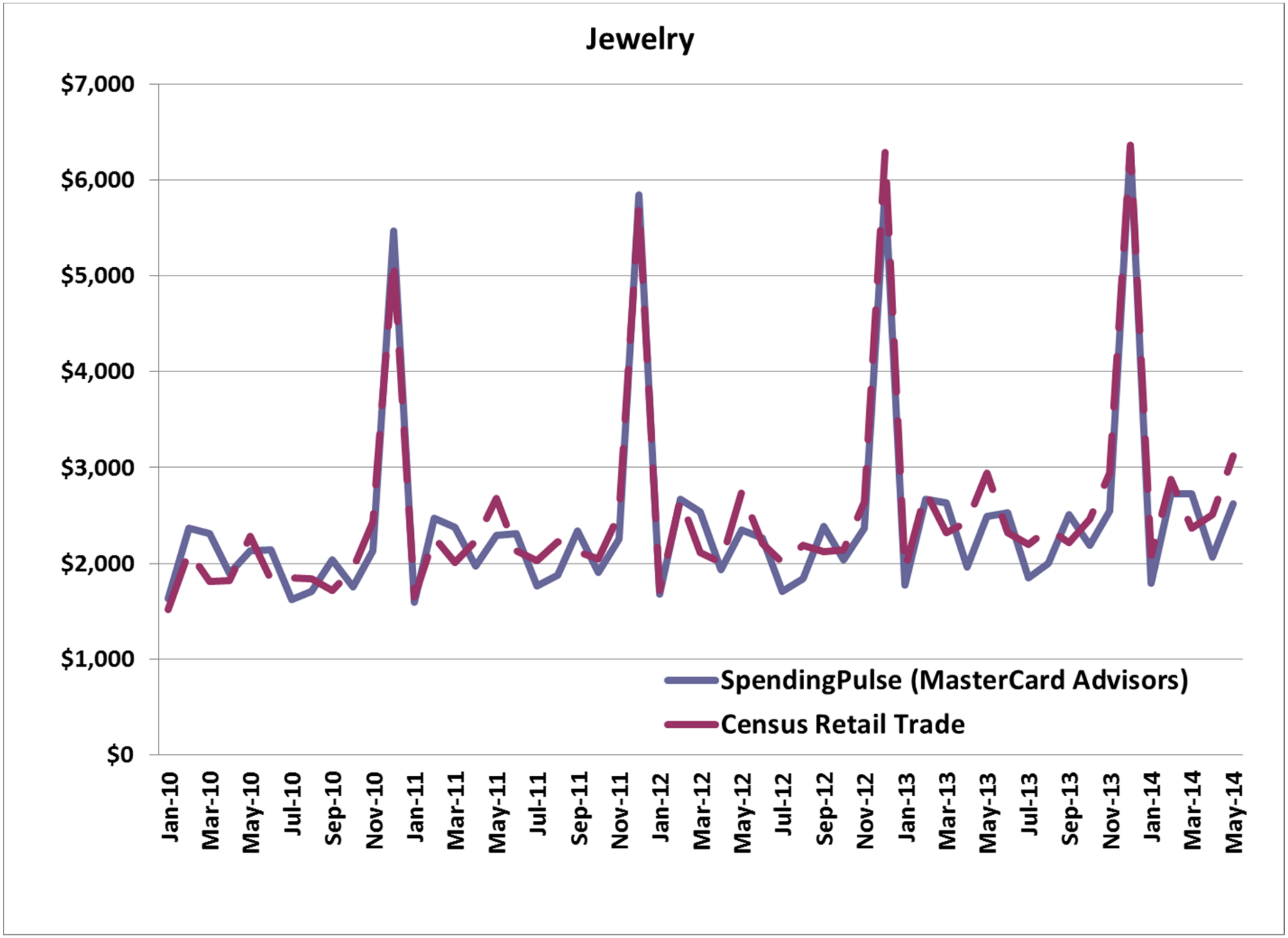
Example: Credit Card Data for Consumer Spending

- Using credit card data collected from the mandatory survey BE-150 to inform its estimates of international travel in the Balance of Payments Accounts
- Exploring use of credit card data to improve estimates of consumer spending, and to develop estimates at the metro area and county levels

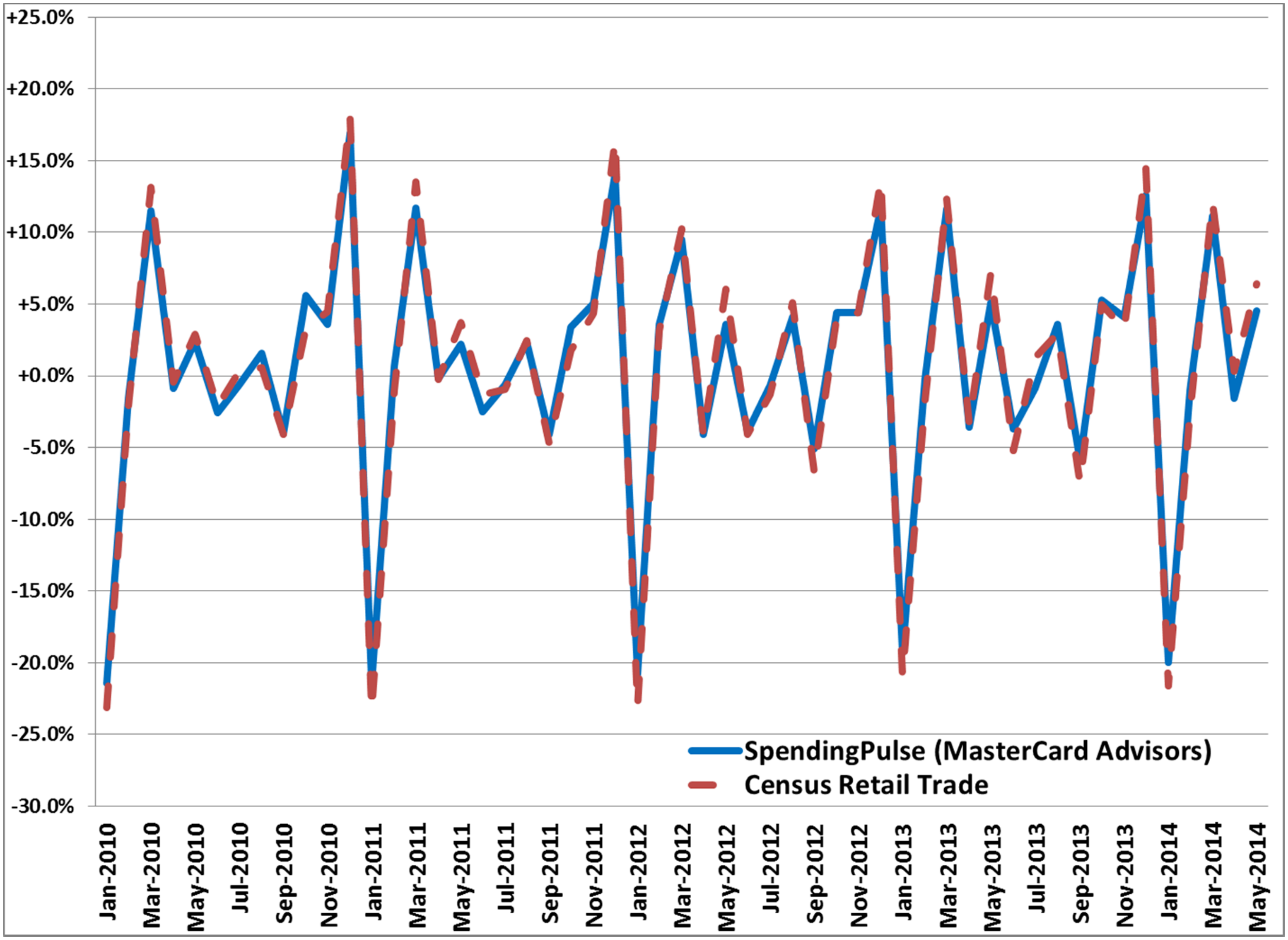
Estimates using Monthly Credit Card Data are similar to Retail Trade aggregates



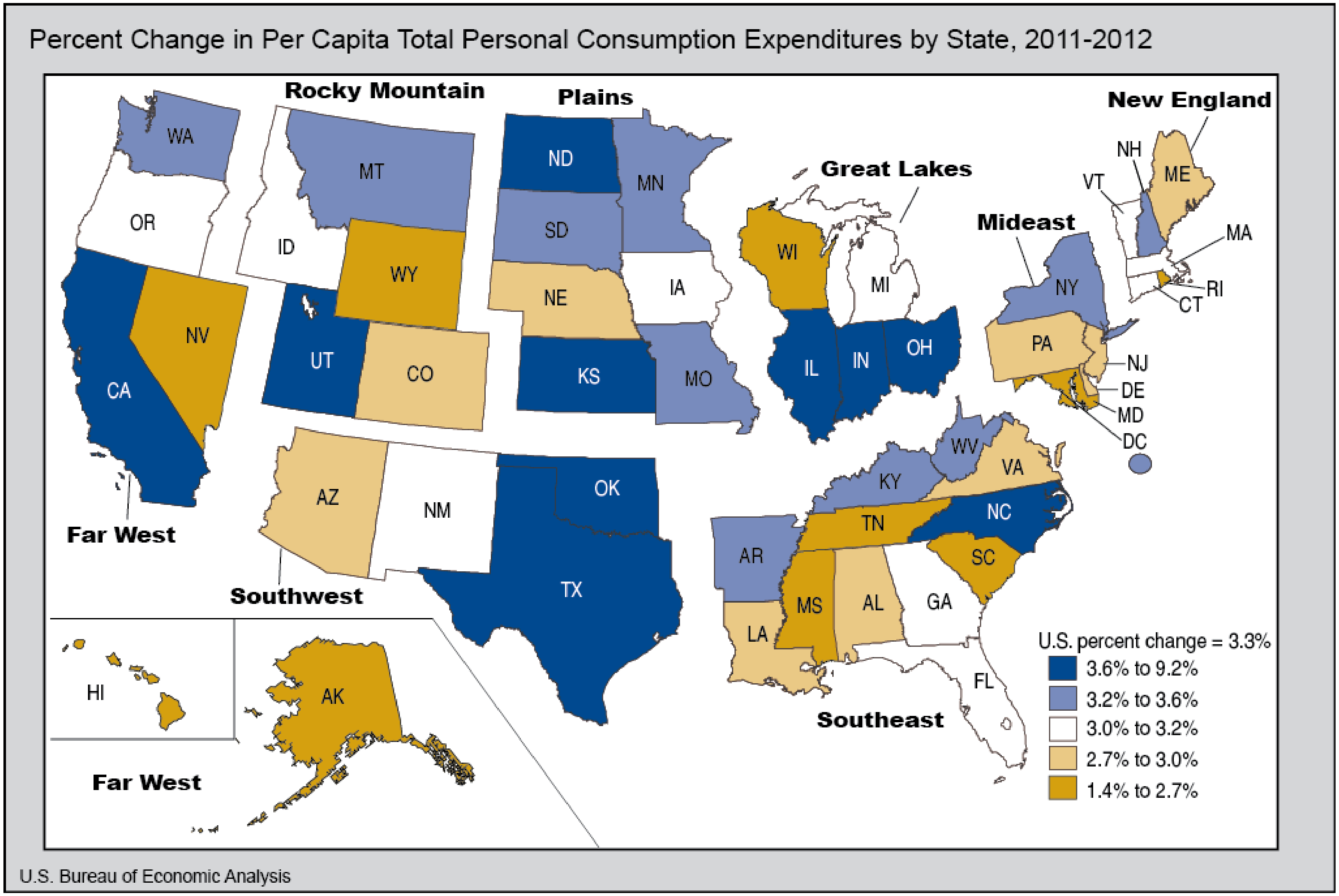
Estimates using Monthly Credit Card Data are similar to Retail Trade aggregates



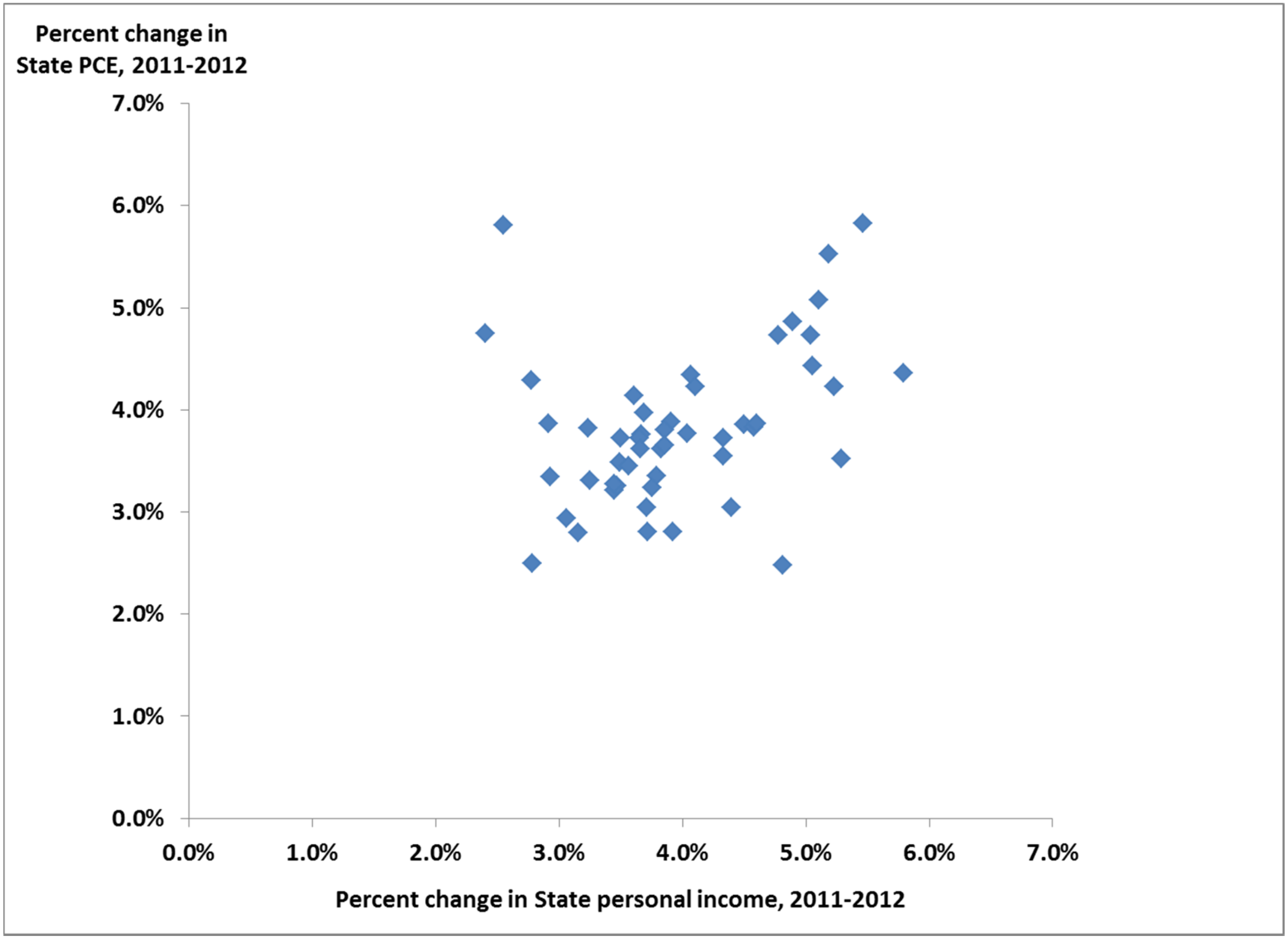
And the monthly changes in SpendingPulse total retail trade (less autos) are similar



Use Commercial data to improve State level estimates of PCE between Economic Censuses



Key is to link change in income to change in spending



Upcoming Meetings on Big Data

- CNSTAT Expert meeting on the use of commercial data in the national accounts, Oct/Nov 2015
  - Presentations by
    - Trivellore Raghunathan (University of Michigan)
    - Simon Wilkie (Microsoft)
    - Jonathan Parker (MIT)
    - Amir Sufi (University of Chicago)
- CNSTAT Panel "Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods"
  - Sponsor - The Laura and John Arnold Foundation
  - Chair – Robert Groves (Georgetown University)
  - Study Director – Brian Harris-Kojetin

Survey Research Methods (2013)
Vol. 7, No. 3, pp. 145-156
ISSN 1864-3361
http://www.surveymethods.org

Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys*

Mick P. Couper
Survey Research Center
University of Michigan

In this paper I review three key technology-related trends: 1) big data, 2) non-probability samples, and 3) mobile data collection. I focus on the implications of these trends for survey research and the research profession. With regard to big data, I review a number of concerns that need to be addressed, and argue for a balanced and careful evaluation of the role that big data can play in the future. I argue that these developments are unlikely to replace transitional survey data collection, but will supplement surveys and expand the range of research methods. I also argue for the need for the survey research profession to adapt to changing circumstances.
Keywords: big data; organic data; social media; mobile surveys; non-probability surveys

> "To everything there is a season, and a time to every purpose under the heaven ... a time to be born, a time to die, a time to plant, and a time to pluck up that which is planted ..." (Ecclesiastes 3:1)

1 Introduction

Has survey research's time come to an end? There are many who suggest that the glory days of surveys are behind us, and we face a future of marginalization if not redundancy (see, e.g., Savage and Burrows, 2007). There are three elements to this. First, with the rise of Big Data[1], when one can collect data on everything that people do, who needs surveys of small subsets of a population? Second, with the rise of opt-in panels, Google Consumer Surveys, Amazon's Mechanical Turk, etc., and other ways to get responses from large numbers of people in relatively little time and at very low cost, who needs probability sample surveys? And third, with the rise of do-it-yourself (DIY) survey tools (e.g., SurveyMonkey), who needs survey professionals? Anyone can do a survey, and – it seems these days – almost everyone does.

Are we redundant? I believe not. In this paper, I review some of the massive changes currently underway in the use of technology – especially social media use and mobile computing – and the implications of these trends on the survey profession. Some take the view that "big data" represents a "brave new world" that will soon replace surveys as the major (or only) source of data on people's attitudes, behaviors, intentions, and the like. This perspective, together with the challenges to traditional surveys in terms of coverage and nonresponse, along with rising costs, may suggest that the survey method has outlived its usefulness. I take a different view, and argue for the important role of surveys – and especially high quality surveys – in our understanding of people and the societies in which we live. I believe that surveys still play a vital role in society, and will continue to make important contributions in the future. However, this does not mean we can be complacent – we do need to adapt as the world around us changes.

It is not my plan to review the technology developments in detail here. This is a well-worn path. There are many who extol the virtues of big data. Similarly, almost every recent presentation on mobile Web surveys reviews all the wonderful things one can do with mobile devices and talks about the rapid penetration of the technology. This was the same kind of excitement that greeted the advent of the Internet, and the development of computer-assisted telephone interviewing (CATI) before that – we are not immune from the hype around new technology. The growth in social media has been similarly well-documented. My goal is to focus not on the technology trends themselves, but on the implications of these trends for the survey profession.

I focus on three key technology-related trends: 1) big data, 2) non-probability samples, and 3) mobile data collec-

Contact information: Mick P. Couper, Survey Research Center, University of Michigan, USA (mcouper@umich.edu)

* Editors' note: This article is not the kind of paper usually published in SRM. However, SRM is the journal of the European Survey Research Association (ESRA), and this text is a reviewed and revised version of the keynote address of the 2013 ESRA conference in Ljubljana. By publishing this article, SRM documents the keynote speech of this important event of its funding institution.

[1] Several others are writing about this topic. For example, Prewitt's (2013) paper appeared as this paper was being completed. In it he talks about the "digital data tsunami" and raises many of the issues regarding big data that are addressed here.

tion. While these are seemingly unrelated, I attempt to show how they raise similar questions for the future of survey research. I discuss each of these in turn before offering some observations on what we can do as survey researchers to respond to the challenge posed by these developments.

2 Big Data

Groves (2011) coined the term "organic data" to describe digital data automatically generated by systems. There are characteristics other than size that describe such data, and "Big Data" (often capitalized) may make one think of "Big Brother," with all the negative connotations[2]. However, "big data" is now part of the modern lexicon, so I will use the two terms interchangeably.

There are three attributes that are generally agreed to describe organic data (see, e.g., Daas, Roos, van de Ven, & Neroni, 2012):

1. volume (exceeds capacity of traditional computing methods to store and process),
2. velocity (streaming data or complex event processing), and
3. variety or variability (raw, messy, unstructured, not ready for processing, does not fit into a relational structure).

In addition to these characteristics of big data, we can identify a number of broad types of organic data, with different implications for access and analysis. These include[3]:

1. Administrative data – data provided by persons or organizations for regulatory or other government activities. Users may assume that the data are confidential and used only for the intended purpose by the agency collecting the data.
2. Transaction data (credit cards, highway/public transport passes, loyalty cards, phone records, browsing behavior, etc.) – data generated as an automatic byproduct of transactions and activities. Users may recognize that the data are being captured and used for the primary purpose of processing the transaction or to facilitate user activities, but may not be aware of secondary uses of the data (e.g., marketing).
3. Social media or social networking data – created by people with the express purpose of sharing with (at least some) others. User expectations about who has access to the data and for what purpose may vary.

Most of my focus is on the second and third types. There are those who argue that with so much data being generated, surveys are no longer of any value. In one provocative view, the title of a 2008 article in *Wired Magazine* posited "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" (Anderson, 2008). In similar vein, Savage and Burrows (2007, 891) argued that, "...where data on whole populations are routinely gathered as a by-product of institutional transactions, the sample survey seems a very poor instrument." In my view, this confuses quantity with quality. My goal here is not to argue for the benefits of big data – I think there are many useful and interesting things that can be done with these data, and they offer exciting opportunities for researchers. My goal is to argue for a balanced view on big data – like all other sources of data (including surveys), organic data have strengths and weaknesses, and understanding these is important to making appropriate use of them.

Some possible limitations of big data – and reasons why I think big data will complement survey data rather than replace surveys – include the following:

Single variable, few covariates

If all we were interested in was temporal trends or geographical variation in a single measure (e.g., the price of fuel, or the spread of influenza), social media analysis or web scraping tools might well give us what we want. But surveys are about much more than estimating a single variable. Social media and transaction data do not have much demographic data. For example, Keeter and Christian (2012) note that demographic information is not available for about 30-40% of Google Consumer Survey (GCS) respondents. For the rest, GCS either uses demographic data provided through Google+ or "assumes" or "imputes" characteristics based on browsing behavior. In their comparison of inferred characteristics from Google Consumer surveys to reported characteristics, Keeter and Christian (2012) found that the inferred gender matches reported gender in about 75% of cases. For age categories, the two match about 44% of the time, but this goes up to 75% of the time when adjacent categories are considered. Similarly, about one-third of Facebook users have no demographic information available (Link, 2013). This limits the kinds of multivariate analyses or subgroup comparisons that are the staple of survey research. Or, as Prewitt (2013) recently put it, big data are case rich but variable poor, while survey data are variable rich but case poor.

Further, the type of data is often limited. Transaction data reveals behaviors – what people are doing, but not why they do it, or what they intend to do in the future. Similarly, social media data might reveal people's thoughts, feelings, preferences, etc., but not necessarily the behaviors that go with those reported views. If one only wanted to measure prices of consumer goods, the Billion Prices Project (see bpp.mit.edu) may give us timely and detailed information. But the Consumer Expenditure Survey (for example; see http://www.bls.gov/cex/) allows us to understand how increases in costs in one type of commodity may result in changes in household expenditures in other areas. For example, do households spend less on medications when food costs go up (or vice versa)? Similarly, we might know that transportation costs are going up, but we don't know whether and how people are changing their travel and other behavior

[2] The recent disclosures about the U.S. National Security Administration's (NSA) use of big data remind one of some of the risks of such data to those who generate it.

[3] There are other types of big data of relevance to national statistics (e.g., passive traffic monitors, movement of goods, environmental monitors). I focus on those involving provision of information by humans.

as a result. We might correlate fuel prices with ridership of public transport or purchase of fuel-efficient cars at an aggregate level using big data, but this may be harder to do at the individual level. For that we need surveys.

Bias

Two types of bias are of concern with regard to organic data. The first is selection bias. Big data tends to focus more on the "haves" and less on the "have-nots". This may also be true of much market research, but social research has traditionally been more interested in the "have-nots". And while almost all of us are users of the new media, we must remind ourselves not to generalize from our own experiences, and remember that while the number of active Facebook users (for example) is enormous, not everyone is on Facebook. Similarly, while almost everyone has heard of Twitter, the number of people who actively tweet is still relatively small (about 13% of the US online population, according to Link, 2013), and highly selective. That is, we should make a distinction between the producers of social media and the consumers of such media. The former may not be representative of the latter, and neither may be representative of the general population. Studying Twitter posts (for example) may be closer to studying elites than the general population. Similarly, there are still sizable minorities of the population who do not use the Internet – thus, for example, those most affected by influenza (the poor, the elderly, the marginalized) may be least likely to search the Internet for help. To the extent that these characteristics are geographically clustered, we may miss key areas affected. Does this invalidate Google Flu Trends as a method of studying the spread of the virus? No, but we must be clear about the selection biases inherent in these kinds of analyses (as with surveys), and understand how they could affect the conclusions we draw.

We also need to understand the limits of transaction data – not everyone uses loyalty cards (for example) or credit or debit cards. Mobile phone (and especially smart phone) penetration is not at 100%. Not everyone communicates by e-mail, and those who do may use different accounts for different purposes. Selection bias can occur at the individual level (e.g., those still using cash) and at the transaction level (e.g., some types of purchases – such as alcohol, cigarettes, condoms, or fast food – may be more likely to be paid for in cash). There are still many ways in which transactions can be conducted without leaving a trace, and many tips and techniques for avoiding being traced (e.g., Singer, 2013). Selection bias is a key feature of organic data (especially of transaction data and social media data, but also of administrative data) and understanding the extent and impact of these biases is a key challenge – and one where we can make a contribution. As with survey data, these biases may be negligible or ignorable in some cases but large and misleading in others.

The second type of bias is measurement bias. Again, this is something that is well known to survey researchers, but has tended to be ignored in the heady rush to exploit the volume of organic data becoming available. Despite the stories one reads about the things people post on Facebook or other social media sites, social media is primarily about impression management (see Boyd & Ellison, 2008). To what extent do people's posts represent their "true" values, beliefs, behaviors, etc.? Similarly, if we counted the number of Facebook friends one has as an indicator of true social network size, we may be seriously wrong. The average Facebook user is estimated to have 229 "friends" (Hampton, Goulet, Rainie, & Purcell, 2011). Again, I'm not saying that Facebook is useless for research purposes, I'm just saying that we need to understand who is using the medium and why they are doing so, in order to understand what biases may exist with social media data.

Volatility or lack of stability

Social media may come and go (remember MySpace? Second Life?), but surveys are relatively constant. This is especially important for trends over long time frames (decades or generations). The rapid rise of Facebook (which was founded in 2004) gets our attention now, but what will Facebook look like five or ten years from now? Will it even exist – what is the half-life of Facebook? Social media may be useful for short-term trends (days or weeks), but may not be stable enough for longer time trends (years, decades). For example, Twitter (which only began in 2006) grew 5000% in the last five years. This means that Twitter today is very different from Twitter five (or even two) years ago. Who knows what Twitter will look like five years from now, or whether it will even still exist? Google itself is just a teenager, with the domain being registered in 1997. There is also rapid evolution in what people share on these sites, and the limits they place on access to their information, especially in response to external events (such as the recent leaks about the US National Security Administration's PRISM surveillance program[4]). As soon as we (research institutions) become interested in a social network or media site, it is probably already past its prime.

Privacy

Related to the issue of volatility is the changing behavior of people using social media and other websites based on concerns about privacy, along with legislation (particularly in Europe) aimed at giving users control over what is collected when they go online. The more the collection and use of big data become broadly known, the more concerned people may become about sharing their information freely – e.g., Wilson, Gosling, and Graham (2012) document some of the changes in Facebook privacy settings over time. This will likely result in an increase in opting out of tracking, rejection of cookies, changes in the amount and type of information shared, use of alternatives to "hide" activities (e.g., paying cash for alcohol and tobacco; using fake e-mail addresses and multiple browsers to confound cookies, etc.), and the development of tools to give users control over what is shared with whom. For example, advertisers have reacted negatively to

[4] http://www.washingtonpost.com/blogs/wonkblog/wp/2013/06/12/heres-everything-we-know-about-prism-to-date/

Microsoft's decision to make the Do Not Track option the default in its new browser (Internet Explorer 10)[5]. Max Frankel (*New York Times*, June 23rd, 2013) noted that: "Privacy is a currency that we all now routinely spend to purchase convenience." But that may not always be the case, and it may not be true of all activities.

Access

Much of the big data being generated is proprietary. It is being used for commercial purposes and has a value (i.e., a price) to those who collect it. Access to data is also restricted for confidentiality purposes, either to protect the identity of participating individuals or to protect the business interests of the entities collecting the data. This means that it may not be freely available – or available at all – to the broader research community. For example, Facebook is not likely to make their database of members available to researchers for sampling or analysis, even at a fee. In addition, the availability of such data may change over time, further adding to concerns about stability. One of the key strengths of surveys, by way of contrast, is public access to the data – conditional on confidentiality restrictions and disclosure limitations. This facilitates reanalysis and replication, which strengthens the underlying value of the data and our faith in the conclusions drawn from the data.

Opportunity for mischief

It is harder to find evidence of this, but I believe that the more people realize that analysis of organic data can influence decision-making, the more likely we are to see attempts to manipulate the system – e.g., to generate interest in a topic or produce the desired results by directly manipulating social media. This is the social media equivalent of ballot-stuffing, which required time and money for call-in polls, but is virtually effortless in the online world, given the ability to write code to generate such content automatically, to create multiple accounts, to generate buzz by re-tweeting, and so on. A story in *The Guardian*[6] in 2011 revealed a US spy operation that manipulated social media, claiming "Military's 'sock puppet' software creates fake online identities to spread pro-American propaganda." Similarly, a recent online story[7] claimed that nearly half of Justin Bieber's 37 million Twitter followers were either fake or inactive. It was recently estimated that about 83 million Facebook accounts (or 8.7% of all accounts) were fake, with 4.8% being "duplicate accounts," 2.4% being "misclassified accounts" (that represent an entity other than the user), and 1.5% being "undesirable accounts" (that purposefully violate Facebook's terms of service, such as spamming)[8]. With increased visibility and importance of big data may come increased attempts to manipulate the data for financial or political gain, or merely to make mischief.

Size is not everything

The characteristic of big data most often mentioned is size. I believe this is the biggest mistake people make with regard to big data. Before we get too excited about the large numbers of people who are using social media, we need to remember that bigger is not necessarily better. Let's take one old example: a sample of 10 million records yielded a response rate of over 23 percent. That's over 2.36 million records – sizeable by any standard. The study was conducted by an organization that had correctly predicted the outcome of 5 previous elections. But the result was a spectacular failure – this is the infamous *Literary Digest* poll of 1936 (see Squire, 1988; Lusinchi, 2012), which called the US election for Landon over Roosevelt. This debacle led to the demise of the *Digest*. Big, but wrong!

On the other hand, it is remarkable that we have to go as far back as 1936 to find such a spectacular failure in election polling. This is an example of cherry-picking that I'll address later – selectively presenting evidence to support arguments against big data. Actually, the 1948 election (Dewey defeats Truman) has been used for decades as an argument for the failure of quota sampling (used by Gallup and all other leading pollsters at the time), and led to the rise of probability sampling. This brings us to the US election of 2012, where Gallup (using probability sampling methods) was one of the furthest from the final outcome[9]. This suggests that all methods need constant evaluation. Election polling (with a few exceptions that would be expected by chance) has had a remarkable run. In several countries, pre-election polls have been used to contest the outcome of elections (i.e., asserting evidence of fraud), suggesting that such polls can at times be even more accurate than a (flawed) count. But it's not just about the size of the sample. And, being accurate once (or even several times) is no guarantee of continued accuracy. This brings me to the final concern about big data.

The file drawer effect

This issue goes well beyond the big data debate, and is worth further attention. The term is attributed to Rosenthal (1979), who wrote: "For any given research area, one cannot tell how many studies have been conducted but never reported. The extreme view of the 'file drawer problem' is that journals are filled with the 5% of the studies that show Type I errors, while the file drawers are filled with the 95% of the studies that show nonsignificant results" (Rosenthal, 1979).

The concern is that much of what we've seen so far is based on selective reporting of findings that support the hypothesis in favor of big data. Aside from the well-known Google Flu Trends (e.g., Dugas et al., 2013), there are many other published papers using Internet searches or Twitter

[5] http://adage.com/article/digital/advertising-week-microsoft-blasted-track/237532/

[6] http://www.guardian.co.uk/technology/2011/mar/17/us-spy-operation-social-networks

[7] http://www.digitalspy.com/music/news/a471915/justin-bieber-twitter-followers-50-percent-are-fake-says-report.html

[8] http://usatoday30.usatoday.com/tech/news/story/2012-08-03/cnbc-facebook-fake-accounts/56759964/1

[9] http://fivethirtyeight.blogs.nytimes.com/2012/11/10/which-polls-fared-best-and-worst-in-the-2012-presidential-race/

analyses to "predict" a variety of things, including voting behavior, problem drinking, mental health, consumer behavior, economic conditions, and the like (see, e.g., Choi & Varian, 2012; Frijters, Johnston, Lordan, & Shields, 2013; Ghosh & Guha, 2013; Lansdall-Welfare, Lampos, & Cristianini, 2012; Paul & Dredze, 2011). While these papers trumpet the success of the method (by showing high correlations between the organic data and benchmark measures), we do not know how many efforts to find such relationships have failed. In one exception, Murphy and colleagues (2011; see also Kim, Hansen, and Murphy, 2012; Kim et al., in press) compared trend analyses regarding the drug *salvia divinorum*, using Twitter feeds and Google search, to data from the National Survey of Drug Use and Health (NSDUH). They find that the trends are quite dissimilar. Specifically, a huge spike in tweets about the drug was associated with a YouTube video of Miley Cyrus smoking *salvia*, without a corresponding change in actual drug use at the time. Similar recent results have been found for Google Flu Trends, with significant errors in both 2009 and 2013 (see Cook, Conrad, Fowlkes, & Mohebbi, 2011; Butler, 2013).

In a humorous example, Leinweber (2007), in a paper originally written in 1995, showed how one can "predict" the S&P 500 index of the US stock market[10] with an R^2 of 0.99 using just three variables: 1) butter production in Bangladesh and the US, 2) cheese production in the US, and 3) sheep production in Bangladesh and the US. The same three variables were useless outside the fitted time period.

The file drawer problem is not limited to new technologies and trends. For example, Hirschhorn and colleagues (2002) conducted a review of 600 positive associations between gene variants and common diseases. Out of 166 reported associations studied 3 or more times, only 6 were replicated consistently. Similarly, Ioannidis (2005) argues that "in modern research, false findings may be the majority or even the vast majority of published research claims" (see also Moonesinghe, Khoury, and Janssens, 2007). In a comparison of publications in 18 empirical areas, Fanelli (2011) found ratios of confirmed hypotheses ranging from 70% (space science) to 92% (psychology and psychiatry). This rate of 92% is far above what should be expected, given typical effect sizes and statistical power of psychological studies (see also Asendorpf et al., 2013; Yong, 2012). This has led some fields to develop ways to encourage the reporting of nonsignificant effects, replications, and the like (see, e.g., http://www.psychfiledrawer.org/).

We as survey researchers are facing a similar dilemma. The papers that "demonstrate" the utility of an exciting new method are more likely to get published than later papers doing the careful but less sexy evaluation of those methods. One simple solution is for journals like *Survey Research Methods* to have a special section for short research notes where such reports are encouraged. Another thing we need is independent evaluations of the trends produced from organic data by those who don't have a vested interest in the outcome – e.g., contrast McDonald, Mohebbi, and Slatkin (2012) with Keeter and Christian (2012). As Carl Sagan and many others have noted, "absence of evidence is not evidence of absence" (Sagan, 1995, 213).

While this is a problem facing all fields of study, because of the large amount of data involved, analysis of organic data may be more likely to yield Type I errors (i.e., finding significant effects or associations where no substantively meaningful effect exists).

My review of concerns about big data may sound like I'm arguing against big data or organic data. This is not the case. I'm convinced that social media research (and big data more generally) has much to offer. The analysis of transaction data is likely to yield many important insights into human behavior that could not be garnered in other ways. Similarly, administrative data have a huge potential. However, I'm equally convinced that these approaches are unlikely to replace survey research. Our role as researchers is to figure out how best to make use of these new opportunities, to expand the range of data we use to understand the societies in which we live. There's a wealth of interesting research opportunities out there for quantitatively-minded researchers. We need to figure out when big data is useful, what biases and flaws may exist, and how we can overcome them. To do this, we need to strip away the hype and examine the evidence in detail – that is, we need to do the research. The same methods and criteria we use for surveys should be useful. As Groves (2011, 869) noted, "The challenge to the survey profession is to discover how to combine designed data with organic data, to produce resources with the most efficient information-to-data ratio." This is where we have important contributions to make.

3 Non-Probability Samples

I will devote less space to the second trend. This is not a new trend, but it is still instructive to review. Non-probability surveys have been around for a long time (see AAPOR, 2013; Baker et al., 2013), but the recent attention that has been paid to such methods can be attributed to the rise of Internet surveys and, more specifically, the development of volunteer opt-in or access panels. Understanding the short history of Web surveys will help us prepare for future technology shifts.

The rise of online opt-in or access panels in the early part of the 21st century was meteoric. Promoters of such panels were claiming that that they make other methods of survey data collection obsolete. One of my favorite quotes from that time is from Gordon Black, then chairman and CEO of Harris Interactive, who stated that "Internet research is a 'replacement technology'—by this I mean any breakthrough invention where the advantages of the new technology are so dramatic as to all but eliminate the traditional technologies it replaces: like the automobile did to the horse and buggy. Market research began with door-to-door household surveys which gave way to telephone polling in the mid-1960s and is now making a quantum leap forward with new Internet research techniques" (Harris Interactive press release, August 1, 1999; see also Couper, 2000).

In the heady early days of Internet panels, the belief was that there was an infinite number of potential survey respon-

[10] http://us.spindices.com/indices/equity/sp-500

dents. It was unthinkable then that the demand for surveys would exceed the supply of respondents. But this is indeed what seems to have happened over the last decade or so. There is increasing evidence that a relatively large number of surveys are completed by a relatively small number of active panelists, many of whom belong to several panels (e.g., Vonk, Willems, & van Ossenbruggen, 2006; Tourangeau, Conrad, & Couper, 2013). The number of surveys requests sent to panelists has sky-rocketed over time. This has led to a rise of concerns about fraudulent or inattentive behavior on the part of panelists, leading some to question the quality of data from such panels (e.g., AAPOR, 2010; Baker et al., 2010). This led the AAPOR Task Force on Online Panels (2010) to conclude that while such panels have a number of uses, "Researchers should avoid nonprobability online panels when one of the research objectives is to accurately estimate population values."

But it's not just the online panels that contributed to this problem. Almost any online transaction these days results in a follow-up satisfaction survey. For those who travel a lot (for example), this can mean several surveys for one trip, including one (or more) for each flight, hotel, rental car, and other activity. Sometimes these surveys take longer to complete than the actual transaction being asked about.

In a way, the very success of Internet surveys has contributed to their possible downfall. There is a parallel to the way the rise of telemarketing affected the telephone survey industry. When something is almost costless and treated like a commodity, it tends to lose value. Beniger was remarkably prescient when he wrote in 1998 about the rise in Web surveys, "Good luck to any serious survey firms which pin much of their futures on the hope of being heard for long above the mounting background noise and confusion of this swelling tide of amateur and slapdash pseudopolls" (Beniger, 1998, 446). Replace "pesudopolls" with "big data analytics" (or any other popular trend) and we see the situation we face today.

Efforts to fix the problems faced by volunteer online panels include a variety of alternative recruitment and selection methods such as river sampling, respondent-driven sampling (RDS), and sample matching (see AAPOR, 2013, for a description of these methods), and the use of Google Consumer Surveys (see, e.g., McDonald, Mohebbi, & Slatkin, 2012; Keeter & Christian, 2012) and Amazon's Mechanical Turk (see, e.g., Berinsky, Huber, & Lenz, 2012). Attention has also focused on improving the design or content of the surveys, with terms like "gamification" and "surveytainment" gaining popularity. In my view, none of these approaches fix the fundamental problem – of demand exceeding supply, of our appetite for data overwhelming the capacity of participants to provide it. Ironically the very success of these panels points to the value of survey data, while at the same time making it harder for everyone to do good surveys because of the saturation problem. For this reason, if the rise in big data means fewer surveys, then maybe this is a good thing. Fewer surveys might mean that those that are done will be of better quality. Scarcity of surveys may also raise their value among potential respondents. It is the ubiquity of surveys and the corresponding commoditization of surveys (Tourangeau, 2010) that have led (in part) to some of the problems we face.

The use of the terms "gamification" and "surveytainment" (see, e.g., Downes-Le Guin, Baker, Mechling, & Ruyle, 2012; Findlay & Alberts, 2011; Puleston, 2011; Tress, Winkler, & Schmidt, 2012) is unfortunate. Trying to turn an otherwise bad survey into a game or a form of entertainment is like putting lipstick on a pig. To be fair, this is not what the proponents of gamification are arguing. Survey engagement (in my view) is a better concept. The idea is not to trivialize the survey enterprise. We want people to take what we do (and what we ask them to do) seriously. Gamifying surveys undermines this and sends a different message. While gamification has been shown to improve a number of metrics such as idea generation, length of open responses, and the like (see Puleston, 2011), this may not be the domain of much standardized survey measurement. However, I believe we should design surveys (both content and presentation) with the goal of fostering user or respondent engagement. This is the basis of user-centered design. We need to see the survey from the respondents' perspective, not our own. In my view, we have become arrogant in our design of surveys, placing increasing demands on respondents, with little thought to their motivation, interest, ability, etc. When concerns are raised, we throw trivial amounts of money at them, in the form of token incentives. I believe we need to meet respondents halfway.

While volunteer online surveys remain enormously valuable and serve many useful purposes[11], they are undergoing a transformation, in part because of the challenges presented by over-saturation, but also in part due to the opportunities presented by big data alternatives. It's going to be interesting to see how this plays out over the next few years.

4 Mobile Data Collection

This brings me to the final technology trend, that of the "mobile revolution." A distinction can be made between three types of mobile use:

1. data collectors (interviewers) using mobile devices (tablets, smartphones, mobile Web) to conduct surveys and collect data,
2. respondents using mobile devices to complete regular Web surveys, and
3. respondents using mobile devices for enhanced data collection (e.g., GPS, photos, ecological momentary assessment (EMA), diary studies, food consumption measures, health monitoring, etc.).

Of most relevance here is the last of these types, but I will indulge in a short detour on the first and a brief comment on the second. The move to tablet-based or hand-held computers finally appears to be here. It has been a long time coming. Based on ergonomic studies conducted in the early 1980s, Statistics Sweden determined the ideal weight of a

[11] To be clear, I have used such panels in much of my own recent work, and think they are an important tool in the survey toolkit.

handheld CAPI computer to be less than 1 Kg (see Lyberg, 1985). In our testing in the US (Couper & Groves, 1992), we came up with a number around 1.6 Kg, which was significantly lighter (by a factor of 5) than all of the available machines at the time. We've had to wait almost 20 years for suitable products to come on the market. The iPad weighs about 0.6 Kg, while the Microsoft Surface is about 0.68 Kg. My point is simply that there are many who criticize the survey profession as being slow to adapt. I believe that, in several instances, the need for the technology is recognized well before such technology is ready for widespread use. Another example is audio-CASI, where the first implementation required interviewers to carry a separate device to generate the sound files because the DOS-based laptop used at the time could not generate sound (see O'Reilly, Hubbard, Lessler, Biemer, & Turner, 1994). It all sounds so quaint looking back, but these were important advances at the time.

Regarding respondents' use of mobile Web, there is a belief (or hope) that mobile Web would bring in different types of people – especially the young, who are currently disproportionately missing from other types of surveys – that is, that technology would compensate for nonresponse bias. So far, the results seem to suggest that we may just be getting more of the same. Those using new technologies to complete our surveys generally seem to be those who would do them anyway using more traditional methods. In this sense, mobile Web may offer more complications to an existing mode rather than solutions to problems we face. But this is an area where further research is needed, and again opportunities abound.

To return to the third type, there are many exciting opportunities for using mobile devices to capture data with greater frequency and fidelity and reduce the need for self-reporting, and there is no shortage of researchers pointing out all the marvelous things that could be done using these devices (e.g., Palmer, Espenshade, Bartumeus, & Chung, 2013). However, to date, almost all of the studies that have demonstrated the use of these devices and apps have been based on volunteers. These volunteers usually have to download and install an app, activate a peripheral device, or otherwise take an active part in collecting the data. These studies have often been restricted to users of particular devices, or to small groups of highly motivated users.

Work on the Dutch LISS panel[12] is one promising exception, and the French ELIPSS panel[13], which is equipping panelists with tablet computers, offers exciting opportunities. But until we can successfully move from small-scale studies of volunteers to implementation among probability-based samples of the general population, these will remain niche technologies (from a general population survey perspective).

Two recent papers from the NTTS conference in Brussels[14] illustrate the challenge. One paper (Biler, Šenk, & Winklerová, 2013) surveyed people in the Czech Republic about their willingness to participate in a travel survey using a GPS device. Only 8% said that they would be willing, while 67% said no (the remainder being uncertain). Another (Armoogum, Roux, & Pham, 2013) asked participants in the 2007-2008 French National Travel Survey about their willingness to accept a GPS device to monitor their travel: 29.8% said yes without condition, 5.1% said yes as long as they could turn it off, and 64.3% said no. Even trained professionals (i.e., interviewers) are not fully compliant – Olson and Wagner (2013) report, for example, that equipping interviewers with GPS-enabled smart phones and having them activate an app to track their work-related travel each day, yielded GPS files for 59.4% of the interviewer-days.

We are all excited about the cool things we as researchers could do with mobile devices, but the question remains, what are people willing and able to do? If we can't answer these questions, we won't be able to defend probability-based surveys against the threat of large data or volunteer surveys. This is one thread that binds these three trends – a point I'll return to later.

Another challenge remains that of coverage. Despite the apparent ubiquity of mobile devices – a recent headline[15] claimed that the number of active mobile phones will exceed the world population by 2014, with more than 100 countries where active cell phones already exceed the countries' population[16] – not everyone has a mobile phone, and not everyone has (or uses) a smartphone. The latest US numbers (June 2013) from the Pew Internet Project[17] suggest that about 91% of telephone-answering adults[18] have a mobile phone, and about 56% have a smartphone. Again, understanding the differences between the "haves" and "have-nots," and what this means for inference to the broader population, is a critical element of good survey research.

Having briefly examined three selected trends driven by technology changes, let me turn to offer a few thoughts on what this all means for the future of surveys, and the future of the survey profession.

5. The Future of Surveys ... and the Surveys of the Future

What ties these three trends (big data, online panels, and mobile data collection together)? While all are at different points in their trajectory, they are all technology trends that have had, or will have, a potentially large impact on the survey profession and the methods we use. In each case, the early proponents of the new methods are (or were) claiming that they will replace "traditional" methods of survey data collection, making current approaches obsolete. At the other end of the spectrum, there are those who bemoan the threat

[12] http://www.lissdata.nl/lissdata/Home

[13] http://www.elipss.fr/elipss/recruitment/

[14] http://www.cros-portal.eu/content/ntts-2013-programme

[15] http://www.digitaltrends.com/mobile/mobile-phone-world-population-2014/

[16] According to the International Telecommunications Union (http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx), there are 126.5 active mobile cellular subscriptions per 100 inhabitants in Europe in 2013 and 109.4 in the Americas.

[17] http://www.pewinternet.org/~/media//Files/Reports/2013/PIP_Smartphone_adoption_2013.pdf

[18] This is a survey conducted by telephone with response rates of 10% for the landline sample and 13% for the cell sample.

that these trends pose for our tried-and-true approaches. My sense is that there are many similarities between these trends – and any future technology trends – that are instructive for the profession.

In 1999, Gordon Black provocatively proclaimed "It's a funny thing about scientific revolutions. People who are defenders of the old paradigm generally don't change. They are just replaced by people who embrace new ideas" (*Wall Street Journal*, April 13, 1999). Are we facing a revolution, a Kuhnian paradigm shift? I'm not convinced that we are. While the changes facing the survey profession are many and large, I don't see surveys going away any time soon. Although surveys will survive as a method of scientific inquiry, we will have to adapt. I see several key areas of adaptation, many of which have been raised by others before. The first two are more specific and practical, while the latter ones are more a call-to-arms or challenge for the profession.

Reducing survey length or burden

We need to match our survey requests to the lifestyles of our potential participants. There has been an increasing disconnect between what we are asking for, and what people may think is reasonable to provide. This is being driven by the rising demand for high-quality survey data, but this is increasingly disconnected from reality. The model where we ask 1-2 hours' worth of questions in a single sitting is no longer sustainable. Who has that kind of time anymore? Or, more critically, those who do are likely to be very different from those who don't.

This model of overly-lengthy surveys is driven by the high cost of asking the first question in interviewer-administered surveys. Given the enormous investment required to find the sample person and get them to agree to do the survey, researchers want to maximize the return on that investment. But this may in fact be counter-productive. It may be – and this is an untested assertion – that a significantly shorter questionnaire may reduce the costs of contacting and persuading sample persons to participate. In part this is the assumption underlying the development of probability-based online panels. We should strive to make the barriers to initial participation low, and build loyalty and commitment over time.

To be provocative, I would go further, and assert that many of the researchers who design the questionnaire would themselves not be willing to do the interview. We have become too removed from the people we are studying. Here's one radical proposal – no one gets to asks questions on a survey unless they themselves will sit down and be interviewed (or complete the questionnaire) as part of the survey pretest. Given the increasing sophistication and complexity of surveys, researchers are increasingly removed from the data collection process, and this has to change. In the team-based approach to survey research, researchers may only feel responsible for a small part of the questionnaire, and may never experience the gestalt of the instrument. Respondents are a precious commodity, a scarce resource, and we should treat them as such. We have to match our survey requests to the lifestyles and expectations of our potential respondents.

How do we find ways to reduce the number of questions we ask? There are several possible approaches, and I think we should be looking into all of them:

1. Work on improving the validity and reliability of single-item measures or short scales, or using item-response theory (IRT) and computerized adaptive testing (CAT) to minimize the number of items asked.
2. Increase the use of data from other sources (whether administrative data or transaction data). We should not ask people to provide answers to questions we can get in other ways – or to questions to which they may not know the answer (see later).
3. Ask less detail, measure with less precision. Our analytic models demand data of increasing fidelity and detail, often exceeding respondents' ability (not to mention willingness) to provide the information. We need more modeling to make estimates based on what we have, rather than increasing our insatiable demand for more data, more variables, and more precision.
4. Make much more use of planned missingness or matrix sampling approaches.

If increased incentives and/or increased effort are the only tools in our toolkit, we are doomed to failure. Until we can also give on the content or length of the survey, we are unlikely to get out of the dilemma we are in. Here I'm a believer in the "less is more" precept of minimalist design popularized by architect Mies van der Rohe.

Using technology

Turning to the second area of adaptation, how can technology help us? We need to think about technology use both by respondents and interviewers. Again, we need to meet respondents halfway, and use the technologies they're already using, and the things they're already sharing, and have them help us. This may mean shorter, repeated measurements rather than single long surveys. Making contact, recruitment, and persuasion are still the key – but we're using old style methods to achieve this at great expense.

Mixed-mode data collection – despite the initial setbacks – is (I believe) still the future of survey research. Responsive or adaptive designs (see, e.g., Groves & Heeringa, 2006; Couper & Wagner, 2011; Schouten, Calinescu, & Luiten, 2011) are gaining ground, but I believe much more could be done. First, we could focus more of our attention on nonresponse bias rather than response rates. But second, we could be thinking about tailored or adaptive designs on a larger scale, including not only mode, incentive, and timing of effort, but also survey content. We need to be more nimble. The era of one-size-fits all approaches may be behind us.

Taken to the extreme, this suggests customized or individualized surveys. We're already doing this with complex computer-assisted interviewing (CAI) instruments with fills, skips, etc., and the increasing use of computerized adaptive testing (CAT), but I'm talking about doing this on a much larger scale. What does this mean for our conception of surveys as standardized measurements on a representative sample of persons? If different subsets of the sample are getting different sets of measures, either based on randomization or on their willingness to participate and provide this information, how do we create rectangular datasets for analysis? In some sense we're already doing this with questions on income (for example). A large number often don't respond, and get followed up with unfolding brackets (which are sometimes themselves not the same for all respondents). With imputation, a single income measure is constructed. This approach also has big technology and process implications – not only for instrument design (CAI programming and testing), but also for documentation and dataset production.

We have to understand how best to do this, and understand what new errors we may be introducing. This is where survey research may be at conflict with itself. One of the fundamental tenets of the survey method is standardization of methods and measurement – everyone is treated the same. In the early days this meant equal probability samples, identical measurement instruments, and standardized interviewing protocols. We have already moved far away from this in terms of sampling – unequal probability samples are now the norm rather than the exception. With the introduction of computer-assisted interviewing (CAI), measurement instruments also became increasingly more customized. Now we use multiple modes of data collection, differential incentives, and a variety of other adaptive approaches. How can we balance the notion of standardization with adaptive and responsive design? This will need good theory, good statistical methods, and good technologies to support.

Understanding the nonresponse problem

This issue has been around since the beginning of surveys, but is increasingly becoming the most pressing issue for probability-based samples. The fundamental problem facing surveys remains that of nonresponse – making contact with people and getting them to respond to surveys. What distinguishes probability-based sample surveys from many other quantitative methods of scientific study (experiments, observational studies, case-control studies, etc.) is that we do not rely primarily on volunteers. But increasingly this is changing, both explicitly (e.g., opt-in or access panels, river samples, etc.) or implicitly (low response rate surveys). What we need is to understand how volunteers differ from non-volunteers on the variables we are interested in measuring and the populations we are interested in studying. This won't be easy, as the very nature of non-volunteers or nonrespondents makes them elusive research subjects. But this is one of the big challenges for survey research in the next decade. There are two related questions we need to try to answer:

1. For probability samples, in what ways are respondents different from nonrespondents, and how this may differ across surveys? This is not just in terms of sociodemographic characteristics (the things we have frame data for, or could correct for), but attitudes, values, behaviors, intentions, etc. More important, we need to answer the question of why they may be different.
2. For non-probability samples and big data analytics, how do volunteers (those who choose to do surveys, sign up for panels, or agree to share their data) differ from non-volunteers?

Tackling these research questions will take new and innovative research methods. Developing theories to explain such differences is the single biggest challenge for surveys. Unlocking this key will help define the role of probability-based surveys for future decades – or lead to the conclusion that probability samples may not be that special after all.

Developing better quality metrics

Next, we need to develop quality metrics to help users differentiate between different types of surveys, or different types of estimates. Unfortunately, the recent work by Groves (2006) and Groves and Peytcheva (2008) makes it clear that this is a hard task. Error – whether sampling or measurement error (as has long been understood), or coverage or nonresponse error (as is only more recently being acknowledged), is a property of a statistic, not of a survey. Replacing response rates with other estimate-level metrics of nonresponse error (for example; see Wagner, 2012) will be a tough sell. But without this, how do we respond to the claims that organic data (and non-probability online surveys) are big, fast, and cheap, and that these factors alone may compensate for lower quality? We can't simply argue that more money means better quality.

The total survey error (TSE) paradigm is a useful framework and a good starting point. But it is rooted in the principles and procedures of probability sampling. We need other ways to quantify the risks of selection bias or non-coverage in big data or non-probability surveys. We need to focus more on costs, not just on errors. TSE remains relevant as an organizing framework but needs to be expanded.

The notion of fitness for purpose has also been around a long time. Quality is not an absolute. It must be evaluated relative to the stated aims of the survey and the purpose to which is put, and the investment (time and money) in obtaining the data. Non-probability surveys and organic data have their place, but so do probability surveys. And we need to develop methods to guide our decisions about which to use when. This is an issue that affects both the producers of data and the consumers of such data, whether analysts or the general public.

Like good wine, the provenance of the data we analyze is important, as is quality. We need to educate users on how to consume data. Sometimes I fear this may be a lost cause. Analytic software makes it too easy for people to conduct analyses without concern for where the data come from or how they are produced. The analytic software we use is agnostic as to the source of the data. Also, the sheer volume of data, and the number of people who directly consume data without regard for source, makes this an almost impossible task. But we must try, at least among ourselves – in the papers we present, in the journal articles we submit and review, in the reports we write. We should take care to point out what we did, and alert readers to the risks of using the data.

Using (and developing) different statistical tools

The kinds of design and analytic problems we are facing require different analytic tools. The methods that many of us learned, which assumed probability-samples with little or no error (other than sampling error) producing rectangular and complete datasets, are increasingly inadequate to handle the complex and messy datasets we now encounter. There's a lot of development already going on in this area, for example, in dealing with missing data, complex hierarchical designs, small area estimation, estimation in the presence of coverage and nonresponse bias, and mixed-mode designs with measurement error (to name but a few). But we also need (for example) new statistical tools to make sense of the masses of messy paradata being generated (see Kreuter, 2013).

On a broader level, we need to be open to other statistical frameworks and approaches to inference, especially for dealing with inference from non-probability based surveys or organic data (see AAPOR, 2013). The probability-based sample survey and frequentist statistical framework are not the only paths to inference. I'm not arguing we should all abandon the frequentist view and become Bayesians (c.f., Little, 2012). But I do agree with Silver (2012, 15) who says "We must become more comfortable with probability and uncertainty. We must think more carefully about the assumptions and beliefs that we bring to a problem." We need tools that match the data we have.

To summarize, I believe surveys will still be around, but they will need to change. We can't cling to the old ways and oppose any new method or approach. Nor can we throw the baby out with the bathwater, and rush to adopt every new method that arises. Big data are here to stay, as are non-probability samples. We have to figure out what method makes sense for which problem.

I find it interesting that those who argue for the superiority of non-probability surveys often use probability-based surveys to demonstrate the quality of their estimates. Similarly, big data estimates are often correlated with survey estimates to evaluate their utility. What would happen if the probability-based surveys were to disappear? We need well-designed and well-executed surveys to serve as benchmarks by which we can evaluate alternative approaches. While high-quality surveys serve this important role of providing a foundation for a vast array of other research, it seems likely that the number and scope of such high-quality benchmark surveys will decline. So far, the demand for all types of surveys – including large-scale, high quality studies like the European Social Survey (ESS), the Survey of Health, Ageing and Retirement in Europe (SHARE), and European Union Statistics on Income and Living Conditions (EU-SILC) – does not seem to have abated, even though there are pressures to do more with less. But I can imagine an effort to consolidate and focus on a few key benchmark surveys while reducing or eliminating overlap or redundancy.

There are lots of interesting opportunities and challenges. Many different skills are needed. We need to set a research agenda that will get us there in the next few years. We're already embarked on this journey, and much good work is already being done in this area. This gives me confidence in the future of our profession.

6 Conclusions

To return to the title of this paper, I don't believe the sky will fall anytime soon. Let me end with two related thoughts. First, a gentle reminder that surveys are tools, and we should not lose focus on the ultimate goals of what we do. Second, I end with some advice for young researchers or those considering getting into this field.

Surveys are a set of tools. More specifically, surveys are a set of tools. There are many different types of surveys and many ways to conduct surveys. So, surveys are like screwdrivers. There are many different types and sizes of screwdrivers, for a range of different purposes. They also vary in quality and cost. But there are also many other tools in a toolbox. Screwdrivers and hammers (for example) serve different functions. Surveys are one of a number of tools we have available for understanding the world around us. They are certainly not the only method, nor are they necessarily always the best. Surveys are particularly good for some things, but not at all good for others.

Sometimes we as survey methodologists fall into the trap of thinking that surveys are the only possible tool. We also get caught up in building the perfect tool, and forget that the tools are not a goal in themselves, but are used for a purpose. Our job is to make better tools, to give the users a range of tools to use in their work, and to guide them in which tool is best for which job. The ultimate goal is to use the tools to make sense of the world around us and, in doing so, help to make a better world.

My view is that we should welcome – rather than fear or oppose – these new developments. They expand the range of tools available to us to understand society. They force is to rethink our assumptions and take a closer look at the methods we're currently using. To continue with the toolkit metaphor, they represent shiny new tools that we can add to our toolkit, enabling us to do things that we couldn't do as well before. But we shouldn't throw away our old tools – and our knowledge of which tools to use for what purpose, and how best to use the tools, remains fundamental. Powerful tools need trained professionals.

Finally, at the risk of sounding arrogant, let me offer some advice for those relatively new to the field. The talk of the obsolescence of surveys may make you wonder what you're getting into. I believe this continues to be a vibrant, rewarding, and fascinating field to work in. There are lots of opportunities to innovate, to develop new methods, and to contribute to our understanding of societies that are rapidly changing. I believe that the training that you have (or are getting) will remain valuable, no matter which direction we take. This will be true even if there are dramatic changes to the way we conduct surveys or measure society. Here are some specific thoughts:

1. Be open to new ideas, but don't be too quick to reject "old" methods. A lot of the theories and methods that have evolved over the decades still apply. One example is the reinvigoration of mail surveys, thought to be near death after the growth of Internet surveys. But there is still clearly a place for mail – at least until the postal service disappears.
2. Look towards the future, but don't ignore the past. It's helpful to remember that the "total survey error paradigm" dates back to the 1940s (Deming, 1944). It's instructive to look back as well as looking forward. Read the old literature – a lot of it is still surprisingly relevant today.
3. Get as much technical and statistical knowledge as you can. Modeling and data analytic skills will always be valuable, I believe. These skills will never be wasted.
4. But don't underestimate the value of good theory. A lot of the issues we face today are crying out for theoretical development – both social science and statistical theories.

For those who are not quite so new, survey research is a dynamic field. Our skills and experience are still relevant today, but are not static. We constantly need to hone our skills, update our knowledge, and expose ourselves to new developments in other disciplines and fields of research and application. This is what makes survey research exciting. While based on strong foundations and a long history of success, survey research is a vibrant, dynamic, and forward-looking field. Long live surveys!

Acknowledgments

This paper was presented as a keynote address at the ESRA conference in Ljubljana in July 2013. I thank Patrick Sturgis and the organizers for giving me the opportunity to present these ideas and encouraging submission of the paper. I am also grateful to Eleanor Singer, Jaak Billiet, and the reviewers for their helpful comments on an earlier draft.

References

AAPOR. (2010). *Report of the AAPOR task force on online panels*. Deerfield, IL: American Association for Public Opinion Research.

AAPOR. (2013). *Report of the AAPOR task force on non-probability sampling*. Deerfield, IL: American Association for Public Opinion Research.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine 16:07*. Retrieved 05/11/2013, from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Armoogum, J., Roux, S., & Pham, T. (2013). *Total nonresponse of a GPS-based travel surveys.* (Paper presented at the conference on New Techniques and Technologies for Statistics, Brussels, March)

Asendorpf, J., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J., Fiedler, K., … Wicherts, J. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108-119.

Baker, R., Blumberg, S., Brick, J., Couper, M., Courtright, M., Dennis, J., … Zahs, D. (2010). Research synthesis: AAPOR report on online panels. *Public Opinion Quarterly*, *74*(4), 711-781.

Baker, R., Brick, J., Bates, N., Battaglia, M., Couper, M., Dever, J., … Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling (with discussion and rejoinder). *Journal of Survey Statistics and Methodology*. doi: 10.1093/jssam/smt008

Beniger, J. (1998). Survey and market research confront their futures on the World Wide Web. *Public Opinion Quarterly*, *62*, 442-452.

Berinsky, A., Huber, G., & Lenz, G. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351-368.

Biler, S., Šenk, P., & Winklerová, L. (2013). *Willingness of individuals to participate in a travel behavior survey using GPS devices.* (Paper presented at the conference on New Techniques and Technologies for Statistics, Brussels, March)

Boyd, D., & Ellison, M. (2008). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, *13*, 210-230.

Butler, D. (2013, February 13). When Google got flu wrong: US outbreak foxes a leading Web-based method for tracking seasonal flu. *Nature*. Retrieved 05/11/2013, from http://www.nature.com/news/when-google-got-flu-wrong-1.12413

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, *88*, 2-9.

Cook, S., Conrad, C., Fowlkes, A., & Mohebbi, M. (2011). Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PLoS ONE*, *6*(8), e23610. doi: 10.1371/journal.pone.0023610

Couper, M. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, *64*(4), 464-494.

Couper, M., & Groves, R. (1992). Interviewer reactions to alternative hardware for computer-assisted personal interviewing. *Journal of Official Statistics*, *8*(2), 201-210.

Couper, M., & Wagner, J. (2011). *Using paradata and responsive design to manage survey nonresponse.* (Invited paper presented at the 58th Congress of the International Statistical Institute, Dublin, Ireland, August)

Daas, P., Roos, M., van de Ven, M., & Neroni, J. (2012). *Twitter as a potential data source for statistics. (Discussion paper 201221)*. Den Haag/Heerlen, The Netherlands: Statistics Netherlands.

Deming, W. (1944). On errors in surveys. *American Sociological Review*, *9*(4), 359-369.

Downes-Le Guin, T., Baker, R., Mechling, J., & Ruyle, E. (2012). Myths and realities of respondent engagement in online surveys. *International Journal of Market Research*, *54*(4), 613-633.

Dugas, A., Jalalpour, M., Gel, Y., Levin, S., Torcaso, F., Igusa, T., & Rothman, R. (2013). Influenza forecasting with Google flu trends. *PLoS ONE*, *8*(2), e56176.

Findlay, K., & Alberts, K. (2011). *Gamification: What it is and what it's not.* (Paper presented at the ESOMAR Congress, Amsterdam, September)

Frijters, P., Johnston, D., Lordan, G., & Shields, M. (2013). Exploring the relationship between macroeconomic conditions and problem drinking as captured by Google searches in the US. *Social Science and Medicine*, *84*, 61-68.

Ghosh, D., & Guha, R. (2013). What are we 'tweeting' about obesity? mapping tweets with topic modeling and geographic information system. *Cartography and Geographic Information Science*, *40*(2), 90-102.

Groves, R. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*(5), 646-675.

Groves, R. (2011). Three eras of survey research. *Public Opinion Quarterly*, *75*(5), 861-871.

Groves, R., & Heeringa, S. (2006). Responsive design for household surveys: Tools for actively controlling survey nonresponse and costs. *Journal of the Royal Statistical Society: Series A*, *169*(3), 439-457.

Groves, R., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly*, *72*(2), 167-189.

Hampton, K., Goulet, L., Rainie, L., & Purcell, K. (2011). *Social networking sites and our lives: How people's trust, personal relationships, and civic and political involvement are connected to their use of social networking sites and other technologies.* Washington, DC: Pew Internet and American Life Project.

Hirschhorn, J., Lohmueller, K., Byrne, E., & Hirschhorn, K. (2002). A comprehensive review of genetic association studies. *Genetics in Medicine*, *4*(2), 45-61.

Ioannidis, J. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), 696-701.

Keeter, S., & Christian, L. (2012). *A comparison of results from surveys by the Pew Research Center and Google Consumer Surveys.* Washington, DC: Pew Research Center for the People & The Press.

Kim, A., Hansen, H., & Murphy, J. (2012). *Methodological considerations in analyzing Twitter data.* (Paper presented at the annual meeting of the American Association for Public Opinion Research, Orlando, FL, May)

Kim, A., Murphy, J., Richards, A., Hansen, H., Powell, R., & Haney, C. (in press). Can tweets replace polls? A U.S. health care reform case study. In C. Hill & E. Dean (Eds.), *Social media, sociality, and survey research* (p. Chapter 3). New York: Wiley.

Kreuter, F. (Ed.). (2013). *Improving surveys with paradata: Analytic uses of process information*. New York: Wiley.

Lansdall-Welfare, T., Lampos, V., & Cristianini, N. (2012). Nowcasting the mood of the nation. *Significance*, *9*, 26-28.

Leinweber, D. (2007). Stupid data miner tricks: Overfitting the S&P 500. *The Journal of Investing*, *16*(1), 15-22.

Link, M. (2013). *Emerging technologies: New opportunities, old challenges.* (Keynote presentation at FedCASIC, Washington, DC, March.)

Little, R. (2012). Calibrated Bayes: An alternative inferential paradigm for official statistics (with discussion and rejoinder). *Journal of Official Statistics*, *28*, 309-372.

Lusinchi, D. (2012). 'President' Landon and the 1936 Literary Digest poll: Were automobile and telephone owners to blame? *Social Science History*, *36*(1), 23-54.

Lyberg, L. (1985). Plans for computer assisted data collection at Statistics Sweden. In International Statistical Institute (Ed.), *Proceedings of the 45th session, book iii* (p. Topic 18.2). The Hague/Leidschenveen: P. Marin.

McDonald, P., Mohebbi, M., & Slatkin, B. (2012). *Comparing Google Consumer Surveys to existing probability and nonprobability based Internet surveys. (White paper).* Mountain View, CA: Google Inc.

Murphy, J., Dean, E., Hill, C., & Richards, A. (2011). *Social media, new technologies and the future of health survey research.* (Paper presented at the Conference on Health Care Research Methods, Peachtree City, GA, April)

Olson, K., & Wagner, J. (2013). *A field experiment using GPS devices to measure interviewer travel behavior.* (Paper presented at the annual conference of the American Association for Public Opinion Research, Boston, May)

O'Reilly, J., Hubbard, M., Lessler, J., Biemer, P., & Turner, C. (1994). Audio and video computer assisted self-interviewing: Preliminary tests of new technologies for data collection. *Journal of Official Statistics*, *10*(2), 197-214.

Palmer, J., Espenshade, T., Bartumeus, F., & Chung, C. (2013). New approaches to human mobility: Using mobile phones for demographic research. *Demography*, *50*(3), 1105-1128.

Paul, M., & Dredze, M. (2011). You are what you tweet: Analyzing Twitter for public health. In *Fifth international aaai conference on weblogs and social media barcelona, july 17-21.* (p. 265-272). Palo Alto, CA: AAAI Publications.

Prewitt, K. (2013). The 2012 Morris Hansen lecture: Thank you Morris, et al., for Westat, et al. *Journal of Official Statistics*, *29*(2), 223-231.

Puleston, J. (2011). Improving online surveys. *International Journal of Market Research*, *53*(4), 557-560.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638-641.

Sagan, C. (1995). *The demon-haunted world*. New York: Ballantine Books.

Savage, M., & Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology*, *41*(5), 885-899.

Schouten, B., Calinescu, M., & Luiten, A. (2011). *Optimizing quality of response through adaptive survey designs. (Discussion paper 201118)*. The Hague/Heerlen: Statistics Netherlands.

Silver, N. (2012). *The signal and the noise: Why most predictions fail – but some don't*. New York: Penguin.

Singer, N. (2013, June 20). Ways to make your online tracks harder to follow. *New York Times, Special Supplement on Big Data*, F4.

Squire, P. (1988). Why the 1936 Literary Digest poll failed. *Public Opinion Quarterly*, *52*(1), 125-133.

Tourangeau, R. (2010). Incentives, falling response rates, and the respondent-researcher relationship. In L. Aday & M. Cynamon (Eds.), *Proceedings of the ninth conference on health survey research methods* (p. 244-253). Hyattsville, MD: National Center for Health Statistics.

Tourangeau, R., Conrad, F., & Couper, M. (2013). *The science of Web surveys*. New York: Oxford University Press.

Tress, F., Winkler, T., & Schmidt, S. (2012). *Surveytainment 2.0: Why investing 10 more minutes in constructing your questionnaire is worth considering.* (Paper presented at the General Online Research conference, Mannheim, March)

Vonk, T., Willems, P., & van Ossenbruggen, R. (2006). *Uitkomsten Nederlands Onlinepanel Vergelijkingsonderzoek (NOPVO).* Amsterdam: Markt Onderzoek Associatie. Retrieved 05/11/2013, from http://www.moaweb.nl/nopvo

Wagner, J. (2012). A comparison of alternative indicators for the risk of nonresponse bias. *Public Opinion Quarterly*, *76*(3), 555-575.

Wilson, R., Gosling, S., & Graham, L. (2012). A review of Facebook research in the social sciences. *Perspectives on Psychological Science*, *7*(3), 203-220.

Yong, E. (2012). Bad copy. *Nature*, *485*(7398), 298-300.

How Falling Gas Prices Fuel the Consumer

Evidence from 25 Million People

October 2015

JPMorgan Chase & Co.
INSTITUTE

About the Institute

The global economy has never been more complex, more interconnected, or faster moving. Yet economists, businesses, nonprofit leaders, and policymakers have lacked access to real-time data and the analytic tools to provide a comprehensive perspective. The results—made painfully clear by the Global Financial Crisis and its aftermath—have been unrealized potential, inequitable growth, and preventable market failures.

The JPMorgan Chase Institute is harnessing the scale and scope of one of the world's leading firms to explain the global economy as it truly exists. Its mission is to help decision-makers—policymakers, businesses, and nonprofit leaders—appreciate the scale, granularity, diversity, and interconnectedness of the global economic system and use better facts, real-time data and thoughtful analysis to make smarter decisions to advance global prosperity. Drawing on JPMorgan Chase's unique proprietary data, expertise, and market access, the Institute develops analyses and insights on the inner workings of the global economy, frames critical problems, and convenes stakeholders and leading thinkers.

The JPMorgan Chase Institute is a global think tank dedicated to delivering data-rich analyses and expert insights for the public good.

Acknowledgments

We thank our research team for their hard work and fabulous contribution to this report, including David Wasser, Pascal Noel and Phoebe Liu.

We would like to acknowledge Jamie Dimon, CEO of JPMorgan Chase & Co., for his vision and leadership in establishing the Institute and enabling the ongoing research agenda. Along with others from across the Firm—notably Peter Scher, Len Laufer, Max Neukirchen, Joyce Chang, Matt Zames, Judy Miller, and Alexis Bataillon—the Institute has had the resources and support to pioneer a new approach to contribute to global economic analysis and insight.

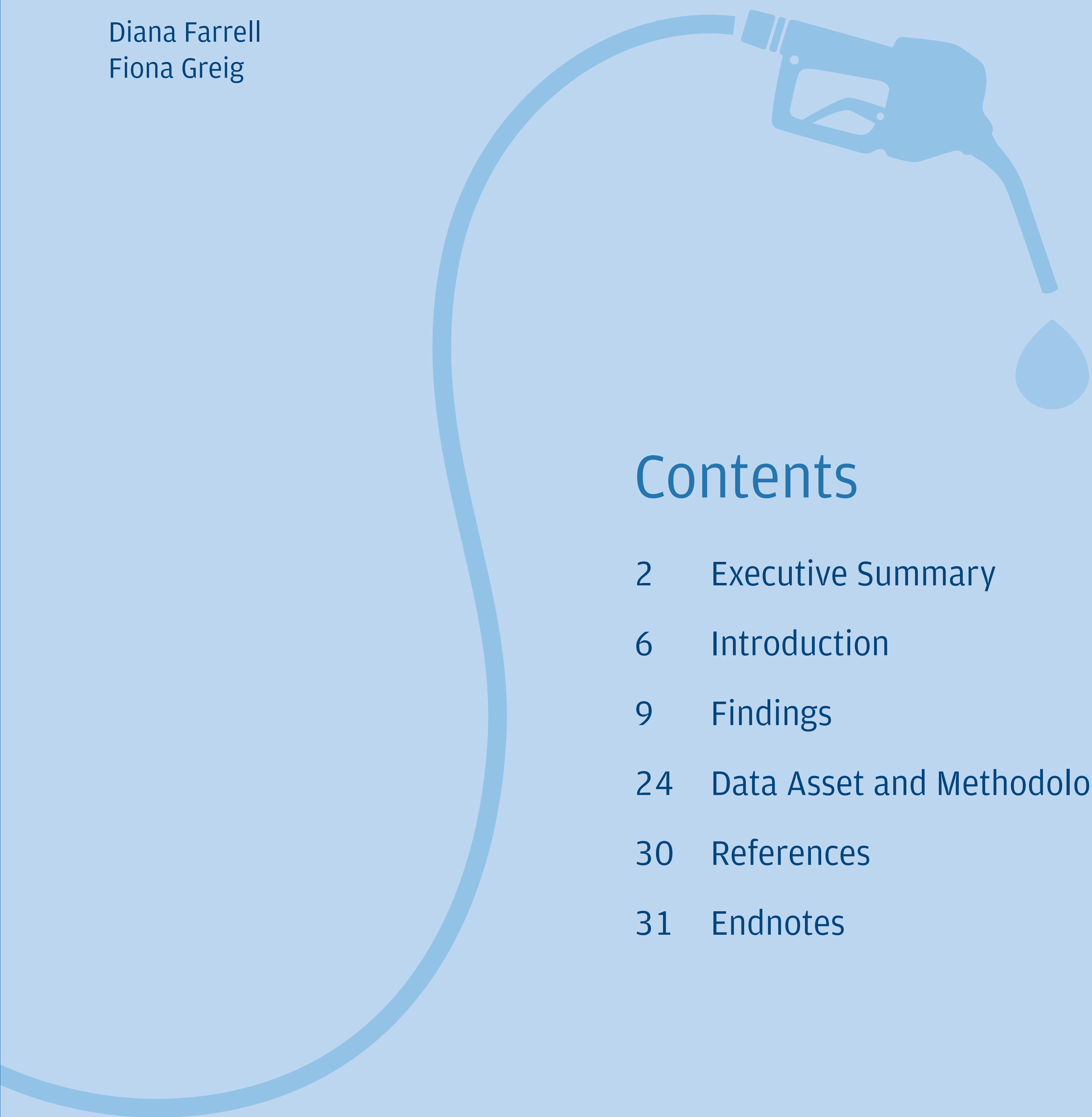
We would also like to acknowledge the contribution of our other researchers, specifically Chris Wheat, Brian Moore and Peter Ganong; and experts within JPMorgan Chase, including Bruce Kasman, Michael Feroli, Jesse Edgerton, Chris Conrad, Tim Ferriter and Scott Prazner. This effort would not have been possible without the critical support of the JPMorgan Chase Intelligent Solutions team of data experts, including Stella Ng, Mohandas Ayikara, Steve Farrell, Joe Bimmerle, Jay Galloway and Michael Solovay, and JPMorgan Chase Institute team members Rachel Pacheco and Kathryn Kulp.

Finally, we would like to acknowledge with gratitude the invaluable input of academic experts who provided thoughtful commentary, including Jim Hamilton, Jonathan Parker and Lutz Kilian. For their generosity of time, insight and support, we are deeply grateful.

How Falling Gas Prices Fuel the Consumer

Evidence from 25 Million People

Diana Farrell
Fiona Greig



Contents

Executive Summary

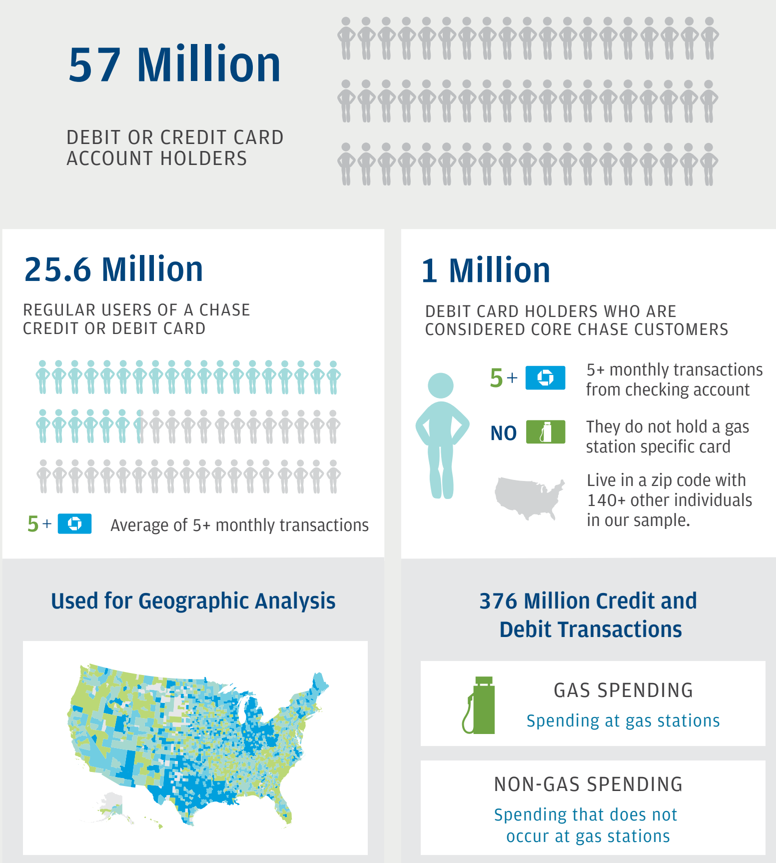
The decline in gas prices since 2014 will save the average American household $700 in 2015 according to U.S. government projections (EIA 2015d). At a time of slow wage growth, this boost in discretionary income is significant. A big question, though, is how individuals are spending that money, if at all.

Until now, the answer to that question has come from surveys or estimates based on aggregate data and has indicated that less than half of the money saved at the pump was spent. However, this report by JPMorgan Chase Institute shows that individuals are spending roughly 80% of that extra money. With lower gas prices expected to last through the year, this extra disposable income is fueling consumer spending on categories other than gas.

Data

From a universe of over 57 million anonymized debit and credit card account holders nationwide, we created samples of 25 million regular card users and 1 million core Chase customers.

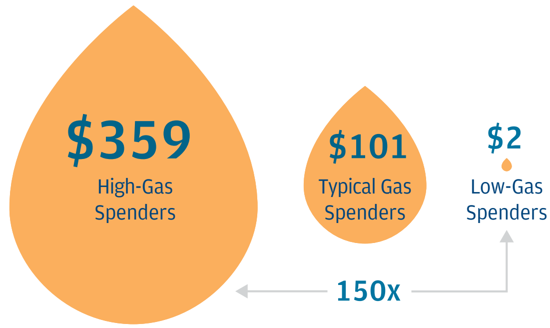
Drawing from a universe of over 57 million anonymized customers, we created samples of 25 million regular debit and credit card holders and 1 million core Chase customers to shed new light on the effects of gas price decreases on consumer spending. We examined spending behavior as prices dropped 45% to their recent trough in January 2015 to determine who experienced the biggest increase in spending power, how much money they spent, and what they bought. Answers to these questions are good indicators of what we can expect going forward if gas prices remain at these lower levels, as projected.



Finding One

Gas spending and the savings associated with gas price declines varied dramatically among U.S. individuals.

Median Americans spent on average $101 per month on gas between December 2013 and February 2014 when gas prices were high. High-gas spenders (the top 20% of gas spenders) spent $359 per month on gas using their credit and debit cards, more than triple the typical American, and low-gas spenders (the bottom 20% of gas spenders) spent only $2 per month, less than 2% of the typical American.

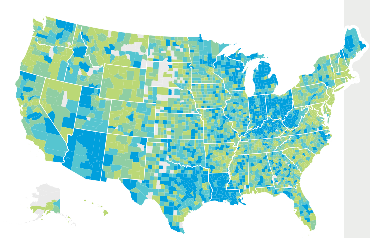


A year later, when gas prices hit their low point, the average American saved $22 per month on gas, but there was significant variation among individuals. Twenty-three percent of the population decreased their gas spending by 50% or more, and 16% increased their gas spending by 50% or more.

Finding Two

People in the South and Midwest spent more on gas and saw larger increases in disposable income when gas prices declined relative to those on the East and West coasts.

People in the South and Midwest spent more on gas and saw larger increases in disposable income when gas prices declined relative to those on the East and West coasts. In the Midwest and South, "higher-impact states," people saw the largest percentage declines in gas prices and gas spending as a fraction of income. In the East and West, "lower-impact states," people saw smaller drops in gas prices and gas spending as a fraction of income. Initially, people in higher-impact states typically paid lower gas prices and consumed more gas than people in lower-impact states.



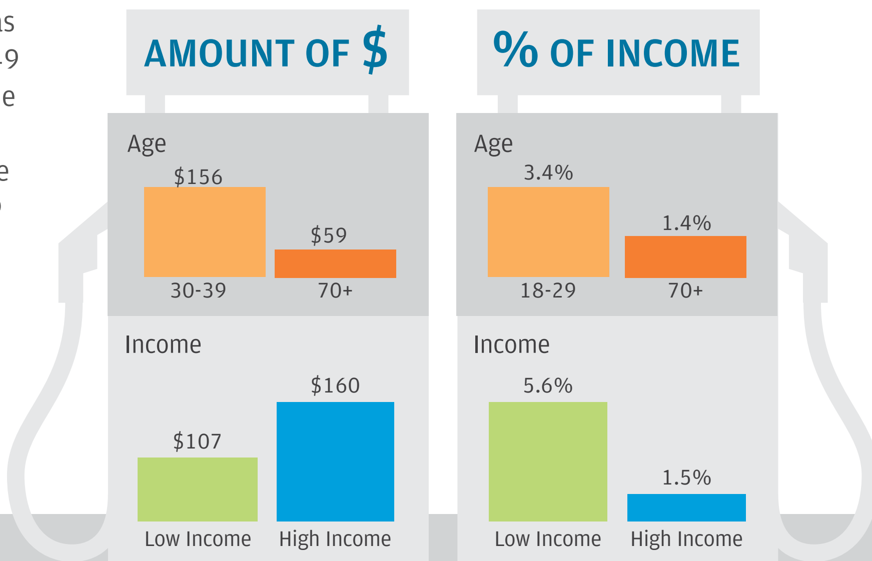
		Initial Gas Prices	Initial Quantity of Gas Consumed	Drop in Gas Prices
Higher-Impact States Large Drop in Gas Spending	**Midwest** IA IN KS MI MO OH SD **South** AL KY LA MS OK TN TX	$		%
Lower-Impact States Small Drop in Gas Spending	**East** CT DE DC FL MA MD NC NY PA **West** AK CA HI NV OR WA	$		%

Finding Three

Savings at the gas pump represented more than 1% of monthly income for low-income individuals and disproportionately impacted younger Americans.

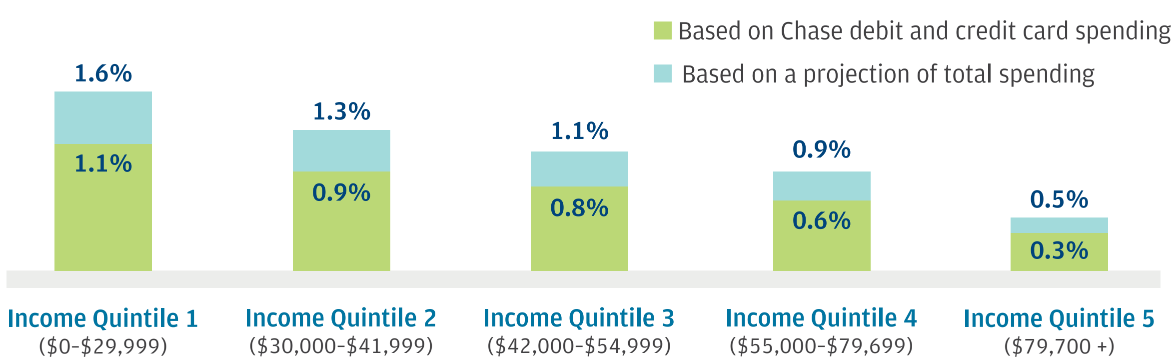
WHO SPENT THE MOST ON GAS?

Although gas spending was highest among men, 30–49 year-olds, and high-income earners, spending on gas represented a larger share of income for men, 18–29 year-olds and low-income earners than for other individuals as a whole.



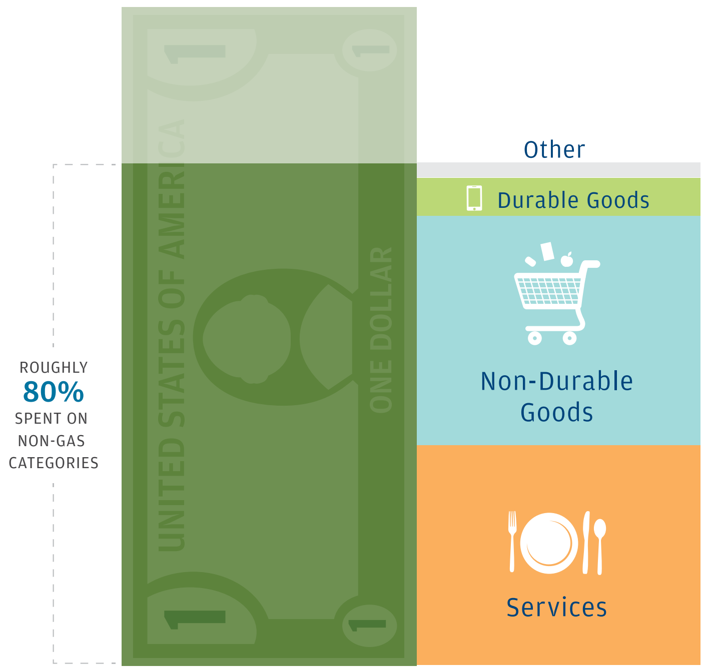
Notably, the recent low point in gas prices in January of 2015 yielded gas savings that represented 1.1% of monthly income for low-income individuals, equivalent to 1.6% of monthly income when projecting total gas spending and not just credit and debit card transactions.

Increase in Purchasing Power from Drop in Gas Spending



Finding Four

Individuals spent roughly 80% of their savings from lower gas prices.



Individuals spent roughly 80% of their savings from lower gas prices. For every dollar less spent at the gas pump, individuals spent roughly 80 cents (72-89 cents) on other things. Almost 20% of the gas savings were spent at restaurants, but department stores, entertainment and electronics and appliances also saw significant gains.

Conclusion

We conclude that people are spending their savings from the pump to a greater extent than previously thought, and that the recent gas price declines are fueling growth in personal consumption in non-gas categories. This boost to consumers spending could be here to stay and even strengthen with time if gas prices remain low or continue to decrease as predicted. On the other hand, a substantial increase in gas prices might proportionately dampen consumer spend. We present evidence that the gains in discretionary spending from lower gas prices disproportionately accrue to low-income individuals, young people, and the Midwest and South, where people tend to spend more on gas. These regional and demographic differences are important inputs as policy makers consider gas tax reforms. Notwithstanding the environmental and infrastructure impacts from increased gas consumption, lower gas prices are good news for the U.S. consumer.

Introduction

The U.S. government projects that American households will save on average $700 this year on gasoline, as the price of a gallon of gas has fallen by nearly $1.50 from its peak of $3.70 in April 2014 and is projected to remain low through 2015 (U.S. EIA 2015d). But who feels the biggest increase in spending power? How much of that extra money do consumers spend, and what do they spend it on?

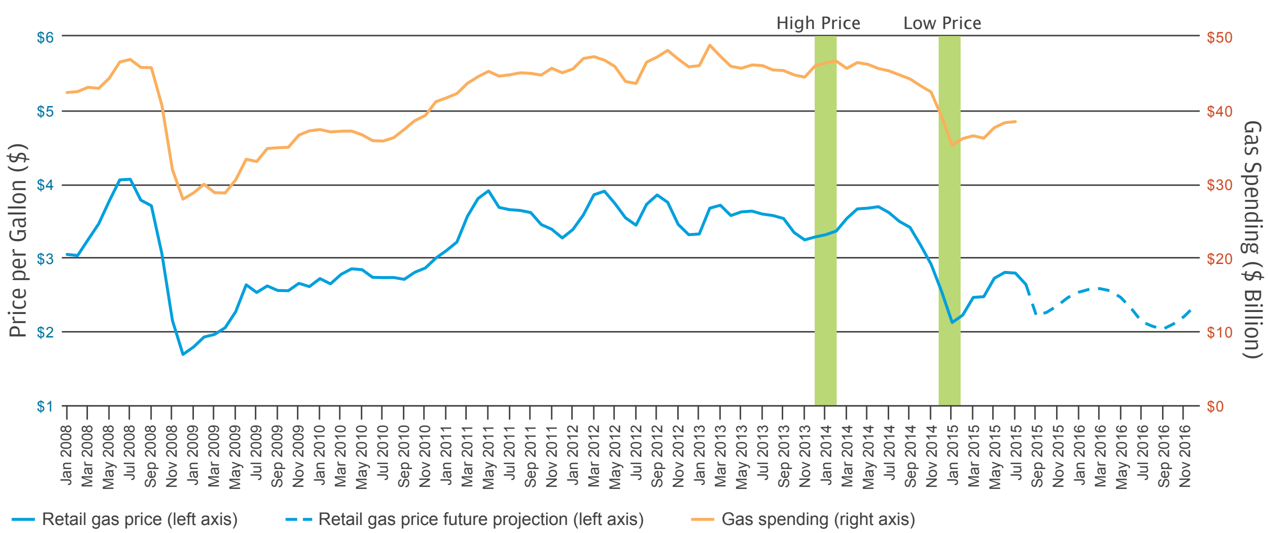
These questions have vexed policymakers and economists in the past, as information regarding the impact of gasoline prices is largely based on consumer surveys and not on actual spending data. The most recent government estimates based on aggregate data comparing the first quarters of 2014 and 2015 suggest that consumers spent only 45 cents for every dollar saved on energy (Furman 2015).

However, research by the JPMorgan Chase Institute shows that consumers spent nearly twice that amount—about 80 cents per dollar saved at the gas pump, and over half of that spending went to restaurants, other services and non-durables. In addition, this research reveals that the Midwest and the South saw the biggest declines in gas spending, primarily because they saw the biggest drops in price and because individuals in these regions consume the most gas. For low-income earners, savings at the pump represented 1.1% of monthly income, equivalent to 1.6% when projecting total spending, or more than half of the growth in income seen by low-income earners between 2013 and 2014. These insights into consumer spending habits shed new light on the effects of price decreases at the gas pump, and help us better understand the role such price declines play in fueling consumer purchases on non-gas categories.

Background

Before we take a closer look at our findings, we review recent trends and projections in gas prices and spending in Figure 1 as reported by the U. S. Energy Information Administration (EIA) and the Census Bureau, respectively. Between April 2014 and January 2015, U.S. gas prices declined 45% from a peak national price of $3.71 per gallon on April 28, 2014, to a low of $2.04 per gallon on January 26, 2015.

Figure 1: National trends in gas prices and spending, with reference to high and low price periods



Source: U.S. Energy Information Administration and U.S. Census Bureau

The last time the U.S. saw a large drop in gas prices was in the last quarter of 2008. Since then gas prices climbed steadily and remained relatively constant between 2012 and the beginning of 2014, fluctuating seasonally between roughly $3.25 and $3.75 per gallon. Prices then fell consistently, dropping by 45% from a peak of $3.71 on April 28, 2014, to a low of $2.04 per gallon on January 26, 2015. Although gas prices have since risen, the EIA forecasts that gas prices will remain below $3.00 through 2015 and 2016.

With the EIA projecting that households will save on average $700 on gasoline in 2015, the gains in disposable income from gas price declines are substantial when compared with recent policy interventions designed to stimulate the economy, as well as ongoing tax policy debates.[1] For example, the Recovery Rebates authorized by the Economic Stimulus Act of 2008 paid between $300 and $600 to each eligible individual. Recent monetary policy interventions generate savings for households in the form of lower interest rates, yielding an annual estimated savings of roughly $600 in lower interest payments on mortgages.[2] The recent fall in gas prices has resulted in a surge in debate and support for gas tax increases.[3]

Low-income individuals experienced the equivalent of a 1.6% increase in income from the fall in gas prices.

As noted above, two critical questions emerge regarding the impact of gas price decreases on the economy: first, who is impacted the most by changes in gas prices; and second, do people spend their savings at the gas pump when gas prices drop, and if so, what do they purchase? In answer to the first question, evidence from the 2014 Consumer Expenditure Survey conducted by the Bureau of Labor Statistics indicates that people spent roughly $205 a month, or roughly 3.7% of their income (before taxes) on gasoline in 2014. Gasoline spending in absolute terms is highest among individuals who are 45–54 years old, live in rural areas and in the West and South regions of the country, and who are high-income earners. However, as a fraction of income, individuals below 30 years old, who live in rural areas and in the Midwest and South, and who are lower-middle income earners (second income quintile) spend the highest fraction of their income on gas.[4] These groups are thus likely to be disproportionately impacted when gas prices fall—they experience the largest increase in discretionary income.

On the second question relating to how consumers react to gas price declines, current research suggests that these price decreases have not generated as much spending as expected. The Council of Economic Advisers estimates that while the recent gas price declines resulted in a 1.1 percentage point decline in energy consumption as a percent of disposable income, this drop only resulted in a 0.5 percentage point increase in non-energy consumption as a percent of disposable income (a 0.6 percentage point increase in personal savings and 0.1 percentage point increase in interest and transfer payments as a share of disposable income) (Furman 2015). This implies a marginal propensity to consume of roughly 45%, much lower than estimates based on past price declines which show that the consumption response exceeds the increase in discretionary income, implying a marginal propensity to consume of greater than 100%.[5]

A recent Gallup poll suggests an even smaller consumer response (Swift, 2015). Although 57% of respondents feel that lower gas prices are making a noticeable difference in their household finances, only 24% say they are spending their gains; the rest are using their gains to pay down bills (42%) or save (28%) (Swift, 2015). Such varying estimates have left policymakers puzzled as to the impact of the recent gas price declines. Federal Reserve Chair Janet Yellen's comments on June 17, 2015, reflect commonly held skepticism on existing data: "I'm not convinced yet by the data that we have seen the kind of response to [the decline in oil prices] that I would ultimately expect. It's hard to know at this point whether or not that reflects a very cautious consumer that is eager to add to savings and to work down borrowing [or that consumers are] not yet confident that the decline in the need to spend [on gasoline] will be permanent" (Federal Reserve Board of Governors, 2015).

It turns out that consumers are spending more of their savings at the pump than has been recently estimated. As described above, existing evidence is suggestive but incomplete. The individual-level surveys have limited sample sizes and are based on self-reported actions rather than economic transactions, and the macro evidence is hard to disentangle from other underlying changes in the economy.

In contrast, the JPMorgan Chase Institute has a rich source of data that offer new, more precise insights into this question. These data include geographically specific, high-frequency, anonymized individual debit and credit card spending from a sample of over 25 million individuals. We analyze these data to describe who is most impacted by gas price changes, and how spending patterns changed after the most recent gas price decline in the second half of 2014.

Our data include debit and credit card spending over the course of 33 months from October 2012 through June 2015. We examine spending during the trough in gas prices, from December 2014 to February 2015, when prices averaged $2.31 per gallon. We compare this spending behavior to one year prior, December 2013–February 2014, when gas prices averaged a dollar higher at $3.31 per gallon. Throughout this paper we will refer to these periods as the High Price period (Dec 2013–Feb 2014) and the Low Price period (Dec 2014–Feb 2015). We chose these periods to maximize the high-to-low variation in gas prices while also allowing us to control for seasonality in gas spending. We explore the impact of gas price declines on consumer behavior across the nation, recognizing the multiple factors that differ by region.

We identify gas and non-gas consumer spending using anonymized data from debit and credit card transactions among Chase customers. We classify all spending at gas stations as gas spending, and spending on everything else as non-gas spending. For most individuals, we know area of residence by zip code, income, age and gender, which allows us to examine consumer behavior across different demographic and geographic groups. Although we do not observe the quantity of gas purchased or the price of gas for each transaction, we use state-specific price data to explore the impacts of gas price declines on the average per capita quantity of gas purchased in each state.[6] The Data Asset and Methodology section provides a more in-depth description of the data and methods used in this report.

Our four key findings are summarized here and described in detail below:

- **Finding 1:** Gas spending and the savings associated with gas price declines varied dramatically among U.S. individuals. The median American spent $101 per month on gas between December 2013 and February 2014 when gas prices were high. High-gas spenders (the top 20% of gas spenders) spent $359 per month on gas using their credit and debit cards, more than triple the typical American, and low-gas spenders (the bottom 20% of gas spenders) spent only $2 per month, less than 2% of the typical American. A year later, when gas prices hit their low point, the average American saved $22 per month on gas, but there was significant variation among individuals. Twenty-three percent of the population decreased their gas spending by more than 50% or more, and 16% increased their gas spending by 50% or more.
- **Finding 2:** People in the South and Midwest spent more on gas and saw larger increases in disposable income when gas prices declined relative to those on the East and West coasts. In the Midwest and South, "higher-impact states," people saw the largest percentage declines in gas prices and gas spending as a fraction of income. In the East and West, "lower-impact states," people saw smaller drops in gas prices and gas spending as a fraction of income. Initially, people in higher-impact states typically paid lower gas prices and consumed more gas than people in lower-impact states.
- **Finding 3:** Savings at the gas pump represented more than 1% of monthly income for low-income individuals and disproportionately impacted younger Americans. Although gas spending was highest among men, 30–49 year-olds, and high-income earners, spending on gas represented a larger share of income for men, 18–29 year-olds and low-income earners than other individuals as a whole. Notably, the recent low point in gas prices in January of 2015 yielded gas savings that represented 1.1% of monthly income for low-income individuals, equivalent to 1.6% of monthly income when projecting total gas spending and not just credit and debit card transactions.
- **Finding 4:** For every dollar less spent at the gas pump, individuals spent roughly 80 cents (72–89 cents) on other things. Almost 20% of the gas savings were spent at restaurants, but department stores, entertainment, electronics and appliances also saw significant gains.

Consumers report that they are using their gains at the pump to pay down debts and save. Our data show they are spending them.

Findings

Finding One

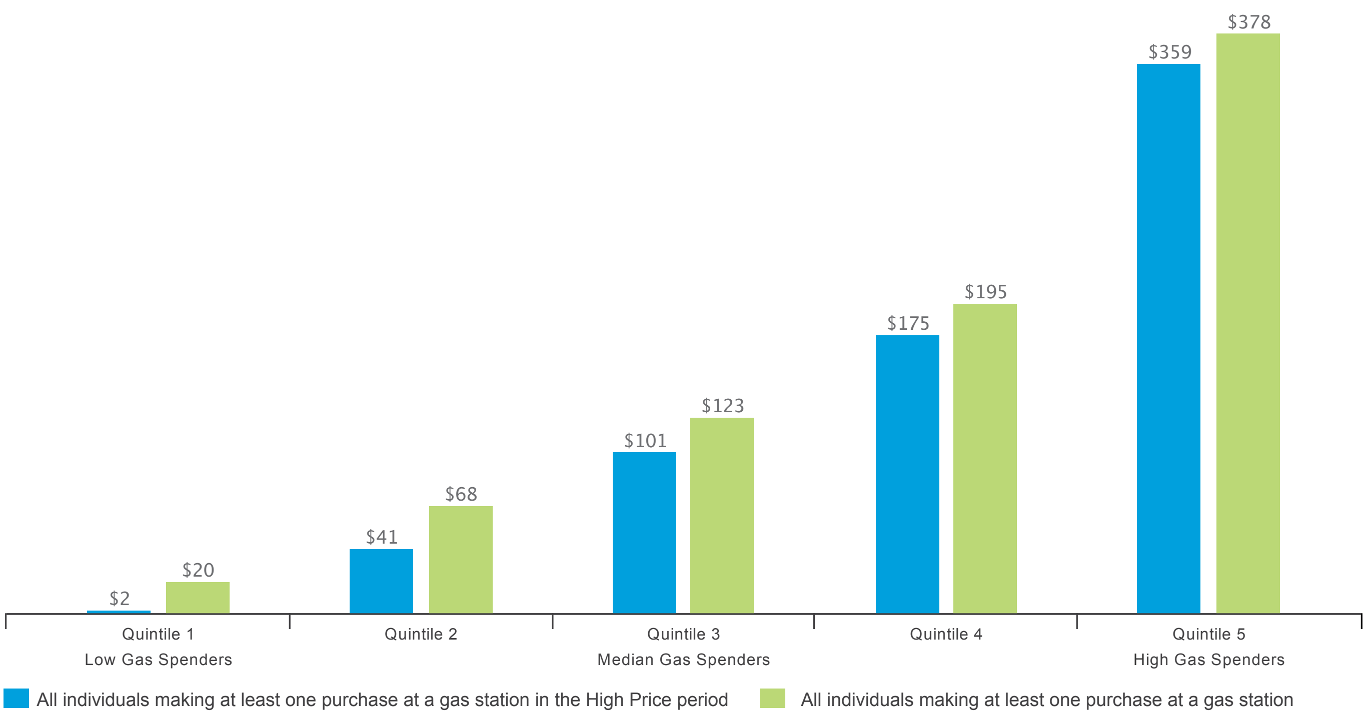
Gas spending and the savings from gas price declines varied dramatically among individuals.

Individual gas spending varies by almost 150-fold between low-gas and high-gas spenders

Before the fall in gas prices, average gas spending for the whole population was $136 per month (in the High Price period). Figure 2 below displays average monthly per person gas spending on Chase credit and debit cards in the High Price period, one year before the national trough in the price of gas. This figure segments the population into quintiles of gas spend. We refer to the bottom quintile as "low-gas spenders," the third quintile as "median-gas spenders," and the top quintile as "high-gas spenders." Median-gas spenders spent on average $101 per month on gas. In contrast, low-gas spenders spent only $2 per month on gas, less than 2% of the typical American, and high-gas spenders spent $359 per month on gas, more than triple the typical American and almost 150-fold that of low-gas spenders.

To ensure that this individual variation is not merely driven by variation in the degree to which people purchase gas using their Chase debit or credit card versus other payment instruments, we calculate this same distribution restricting our sample to the 78% of people who show any gas purchase in the High Price period. As shown in Figure 2 below, within this subsample, we still see almost a 20-fold gap between low-gas and high-gas spenders.[7]

Figure 2: Average monthly gas spending by quintile of gas spend, High Price (Dec 2013–Feb 2014)



Source: JPMorgan Chase Institute

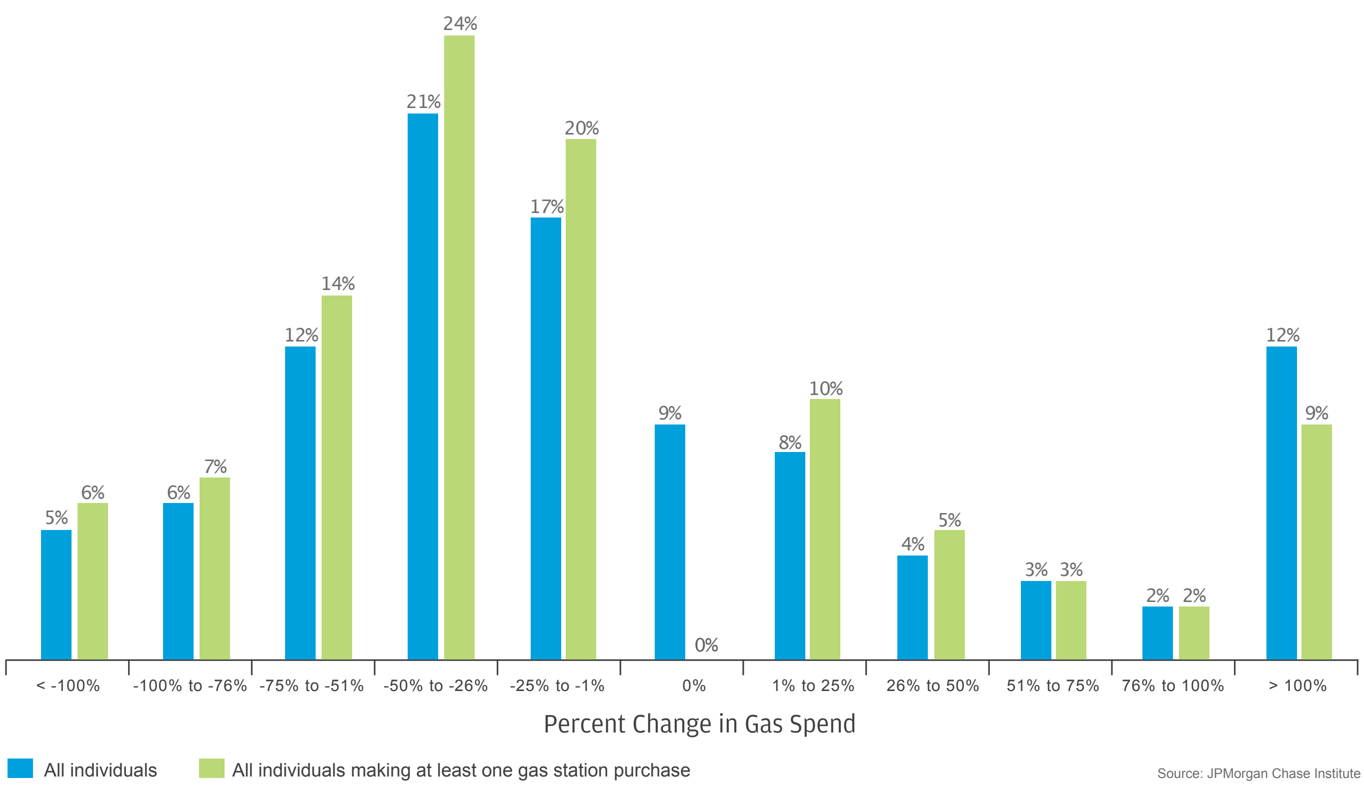
The levels of monthly gas spending that we observe are significantly lower in the Chase sample for 2014 ($146) than the $206 reported in the 2014 Consumer Expenditure Survey. This gap likely exists because people may pay for some of their gasoline using cash, check or a non-Chase card. In addition, the gap could be partially due to differences between the Chase sample and the nation. This gap will also explain why, for 2014, we estimate that individuals spent only 2.9% of their income on gas compared to the national average of 3.7% of income, according to the Consumer Expenditure Survey.[8]

High-gas spenders spend $359 per month on gas, more than triple the typical American and almost 150-fold that of low-gas spenders.

Between the High Price period and the Low Price period, the average person saw a $22 decrease in gas spending, but there was again significant variation among individuals. As shown in Figure 3 below, 62% of all individuals decreased their gas spending, including 23% of people who decreased their gas spending by 50% or more. Nine percent of people saw no change in their gas spending, and 29% increased their gas spending, including 16% of people who increased their gas spending by more than 50%. Again, we find similar degrees of variation in the change in spending on gas when we examine only individuals who showed any gas spending in the High Price period. Among this sample, 71% of people spent less on gas in the Low Price period compared to the High Price period, and 27% decreased their gas spending by more than 50%. The remaining 29% increased their gas spending, including 14% who increased expenditures on gas by more than 50%.

Taken together, Figures 2 and 3 convey not only the variation among individuals in terms of gas spending but the degree of volatility of gas spending due in part to the gas price decline. Next we explore regional and demographic differences in levels and changes in gas spending.

Figure 3: Distribution of the percent change in monthly gas spending between the High Price period and the Low Price period



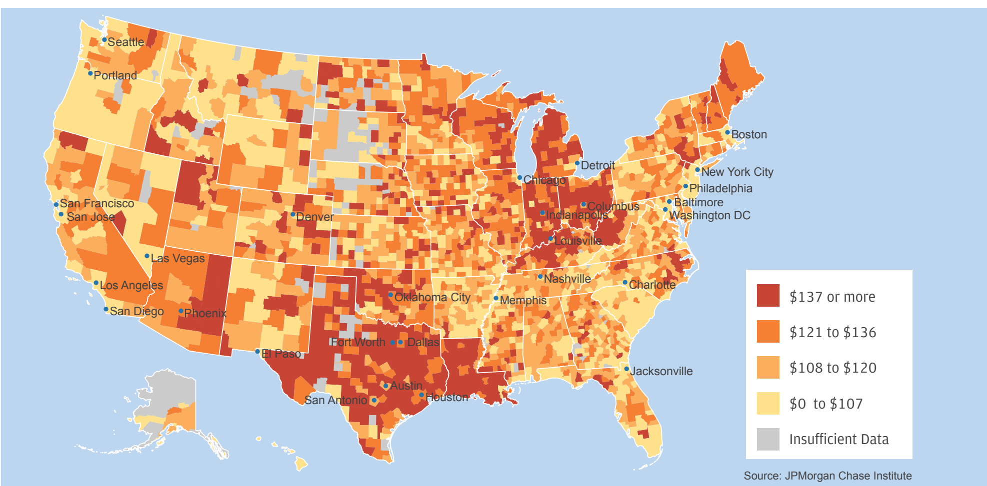
Finding Two

People in the South and Midwest spent more on gas and saw larger increases in disposable income when gas prices declined relative to those on the East and West coasts.

People in the Midwest and South spent the most on gas

Gas spending varies tremendously by geography. Although Chase's branch footprint covers 23 states, we are able to observe spending behavior across the nation by aggregating observed spending on both debit and credit cards. As described in the Data and Methodology section, the mix of debit and credit holders in any given state varies significantly.[9] Figure 4 maps the different levels of observed average monthly gas spend by county in the High Price period, based on 25.6 million frequent Chase credit or debit card users. Figure 5 shows observed spend levels as a percent of income. Evident from these maps is that people in the Midwest and the South spent the most on gas both in absolute terms and as a fraction of income. In addition, the brighter spots around major cities indicate that people in urban areas spent less on gas, particularly as a fraction of their income.[10]

Figure 4: Average gas spend by county in the High Price period (Dec 2013–Feb 2014)



Top 5 Counties*

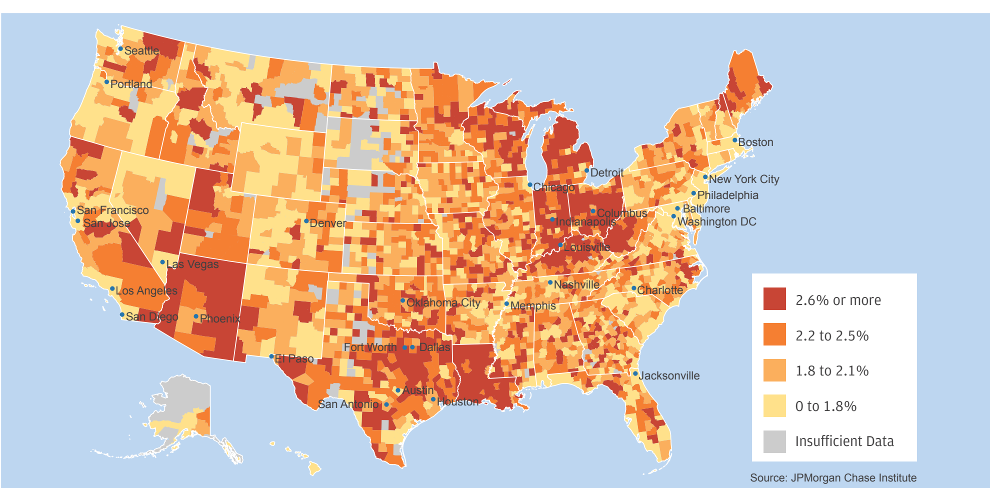
Allen, Louisiana
Armstrong, Texas
Cameron, Louisiana
Greenlee, Arizona
La Salle, Texas

Bottom 5 Counties*

Bronx, New York
District of Columbia
Hudson, New Jersey
Kings, New York
New York, New York

* Counties within each list are displayed in alphabetical order

Figure 5: Average gas spend as a percent of income by county in the High Price period (Dec 2013–Feb 2014)



Top 5 Counties*

Clay, West Virginia
Crawford, Indiana
Greenlee, Arizona
Lincoln, West Virginia
Morrow, Ohio

Bottom 5 Counties*

District of Columbia
Hudson, New Jersey
Maui, Hawaii
New York, New York
San Francisco, California

* Counties within each list are displayed in alphabetical order

We further explore this geographic variation by ranking states according to gas spend, both in levels and as a percent of income, in Figure 6 below. Individuals in West Virginia spent the most on gas both in absolute terms and as a fraction of their income. The top 10 states in terms of gas spending were all in the South (Kentucky, Louisiana, Oklahoma, Texas, West Virginia) and the Midwest (Indiana, Ohio, Wisconsin), with the exception of New Hampshire and Maine. The top 10 states in terms of spending as a percent of income were again all in the South (Kentucky, Louisiana, Oklahoma, Texas, West Virginia) and the Midwest (Indiana, Michigan, Ohio, Wisconsin), with the exception of Arizona.

The bottom 10 states in terms of spending on gas are all in the East (District of Columbia, Maryland, New Jersey, New York, South Carolina, Virginia) and the West (Alaska, Hawaii, Montana, Oregon). Individuals in the bottom 10 states in terms of percent of income spent on gas are similarly mostly in the East (New Jersey, New York, Massachusetts, Connecticut, District of Columbia, Maryland) and the West (Alaska, Hawaii, New Mexico), with the exception of Virginia and South Carolina.

Figure 6: Average gas spend by state in the High Price period (Dec 2013–Feb 2014)

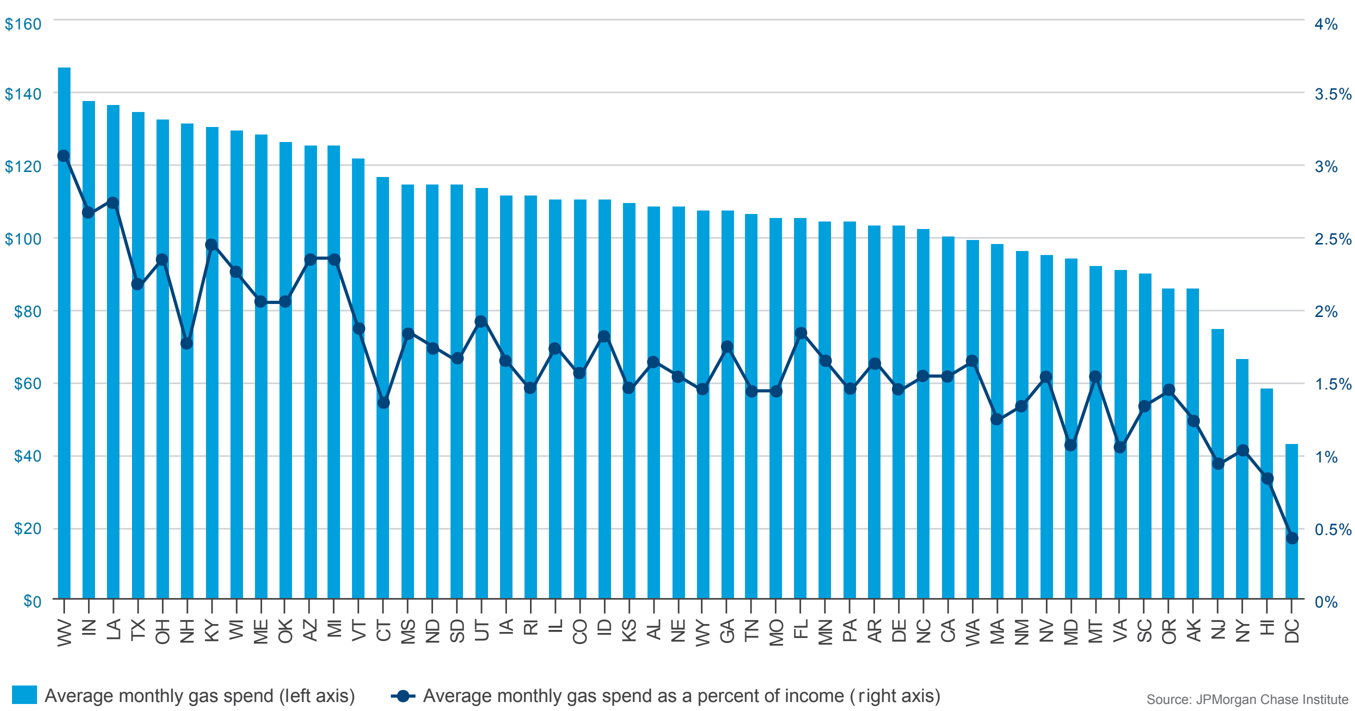
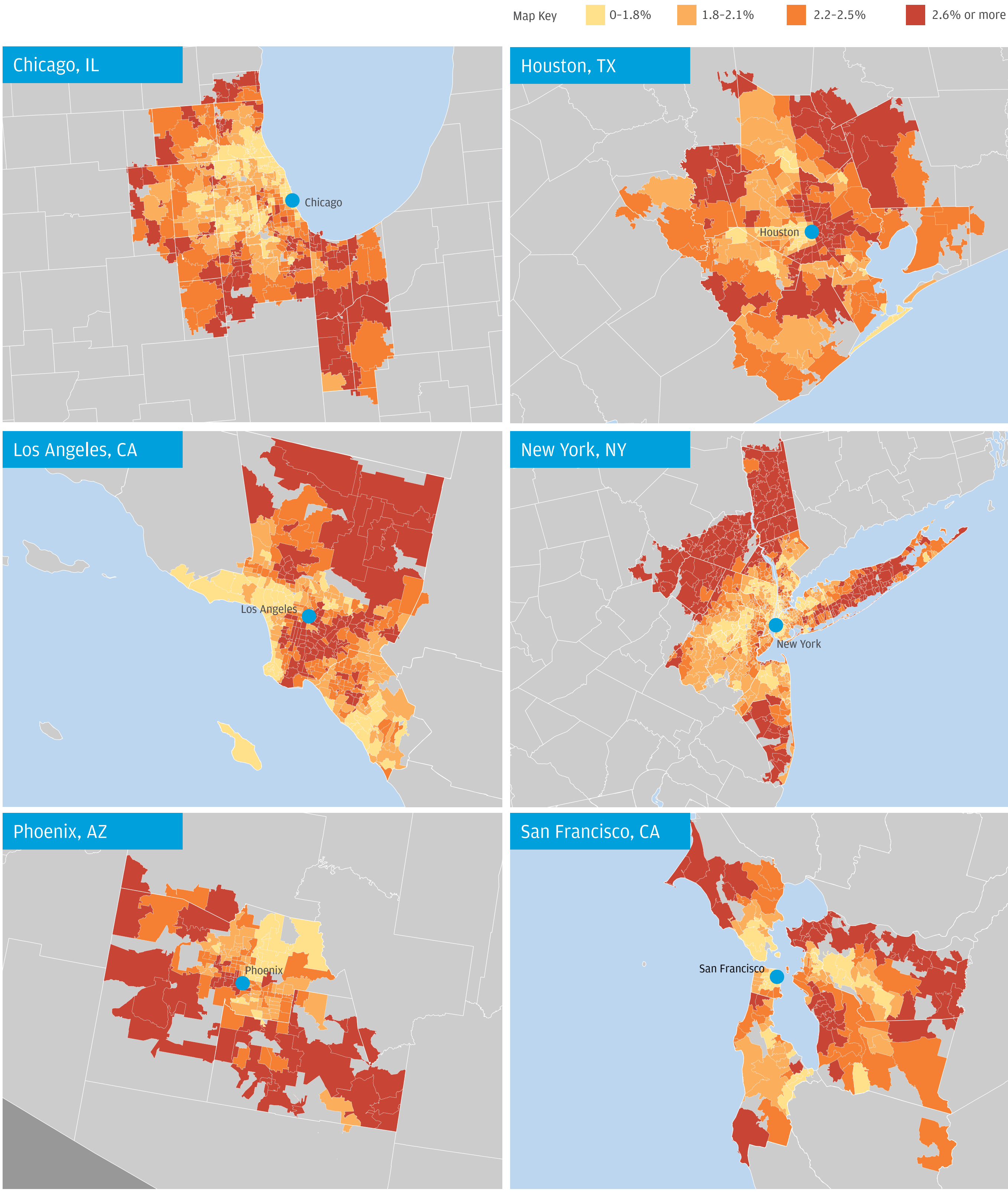


Figure 7 zooms in on six major cities—Chicago, Houston, Los Angeles, New York, Phoenix and San Francisco—and displays average monthly spending on gas as a percent of income by zip code. Evident from these more granular maps is that gas spending as a fraction of income is significantly higher in suburban as well as lower-income areas within metropolitan areas.

Gas spending as a fraction of income is significantly higher in suburban areas surrounding inner-city cores.

The recent gas price declines put more discretionary income into the pockets of people in the Midwest than anywhere else.

Figure 7: Average gas spend as a fraction of income by zip code within 6 metropolitan areas in the High Price period (Dec 201–Feb 2014)

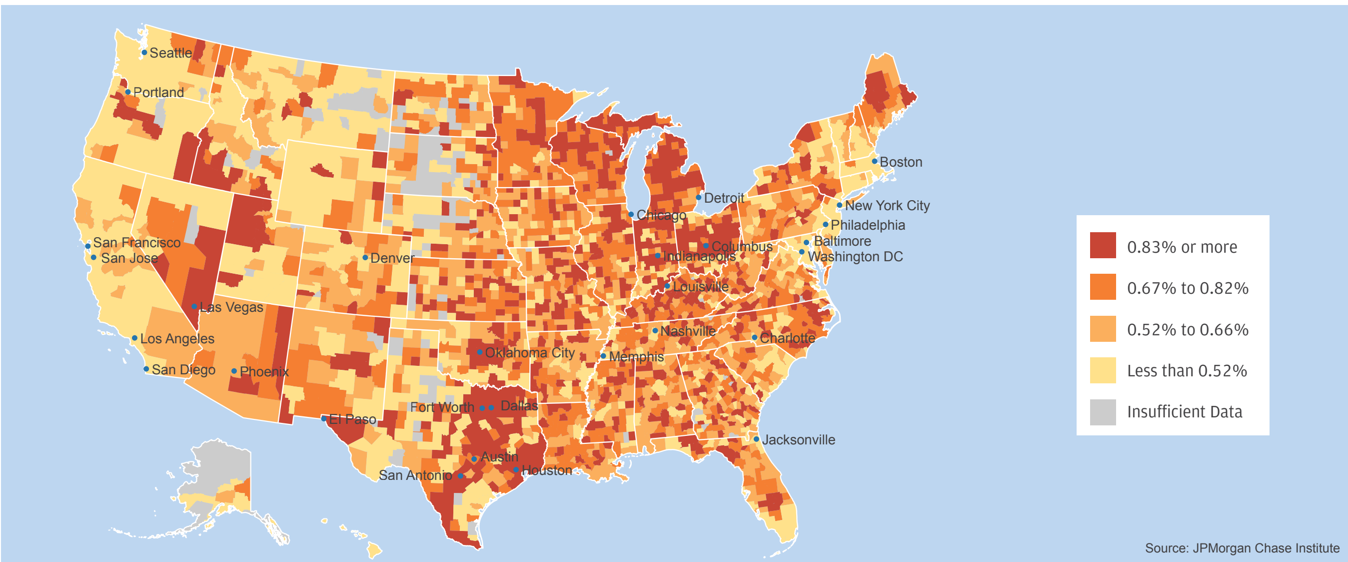


Source: JPMorgan Chase Institute

The Midwest saw the largest drops in gas spending

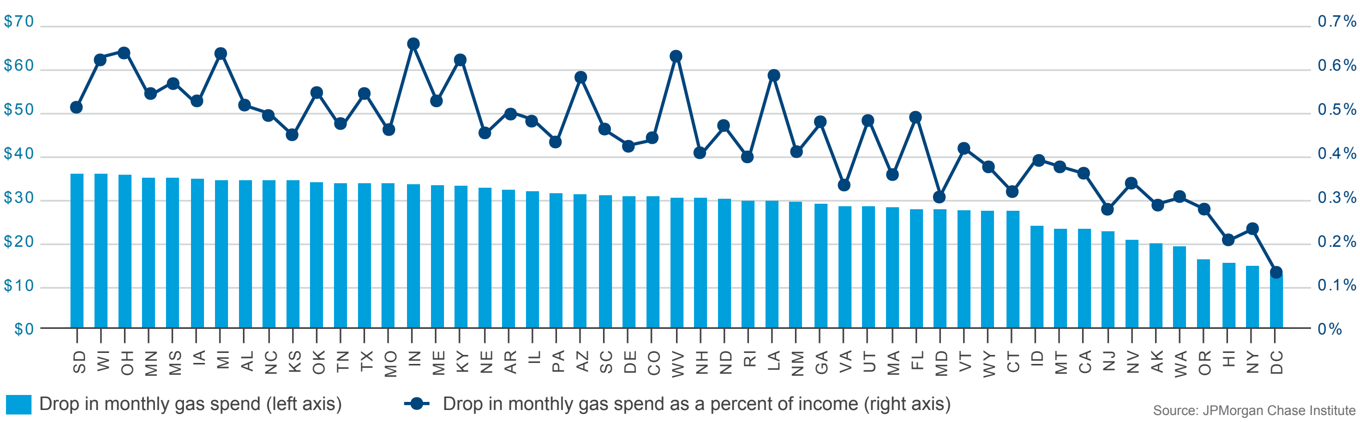
Having explored how gas spending varies by region, we now examine which areas of the country saw the largest drop in gas spending and the largest equivalent increase in disposable income as a result of the gas price declines. The map below shows the change in gas spending between the High Price and Low Price periods as a fraction of income. Counties that saw the highest increases in disposable income from lower spending on gas are concentrated within the Midwest and southern Plains states.

Figure 8: Drop in average monthly gas spending as a percent of income by county between the High Price period (Dec 2013–Feb 2014) and the Low Price period (Dec 2014–Feb 2015)



The recent gas price declines put more discretionary income into the pockets of people in the Midwest than anywhere else. As shown in Figure 9, the 10 states that saw the largest drops in gas spending are all in the Midwest (Iowa, Kansas, Michigan, Minnesota, Ohio, South Dakota, Wisconsin) and the South (Alabama, Mississippi, North Carolina). In terms of spending on gas as a fraction of income, the top 10 states are again all in the South (Kentucky, Louisiana, Mississippi, Oklahoma, West Virginia) and the Midwest (Indiana, Michigan, Ohio, Wisconsin). The 10 states that saw the smallest drop in gas spending are all in the East (District of Columbia, New Jersey, New York) or the West (Alaska, California, Hawaii, Montana, Oregon, Washington). Gas spending as a fraction of income fell the least in many of these same states (Connecticut, District of Columbia, Maryland, New Jersey, New York, and Virginia in the East; and Alaska, Hawaii, Oregon, and Washington in the West).

Figure 9: Drop in gas spending between the High Price period (Dec 2013–Feb 2014) and the Low Price period (Dec 2014–Feb 2015), by state



Overall, Midwestern and Southern states were "higher-impact" states while Eastern and Western states were "lower-impact" states

In any given state, the change in gas spending is determined by the initial price of gas and quantity of gasoline purchased in the High Price period as well as the changes in price and quantity purchased after gas prices dropped. These four factors are interrelated. We observe spending on gas, which we combine with state-level gas prices in order to infer average monthly quantity of gas consumed by each state.[11] We find several important geographic differences, both in levels and in changes in gas prices and quantity consumed, which underpin differences in gas spending across states.

2x

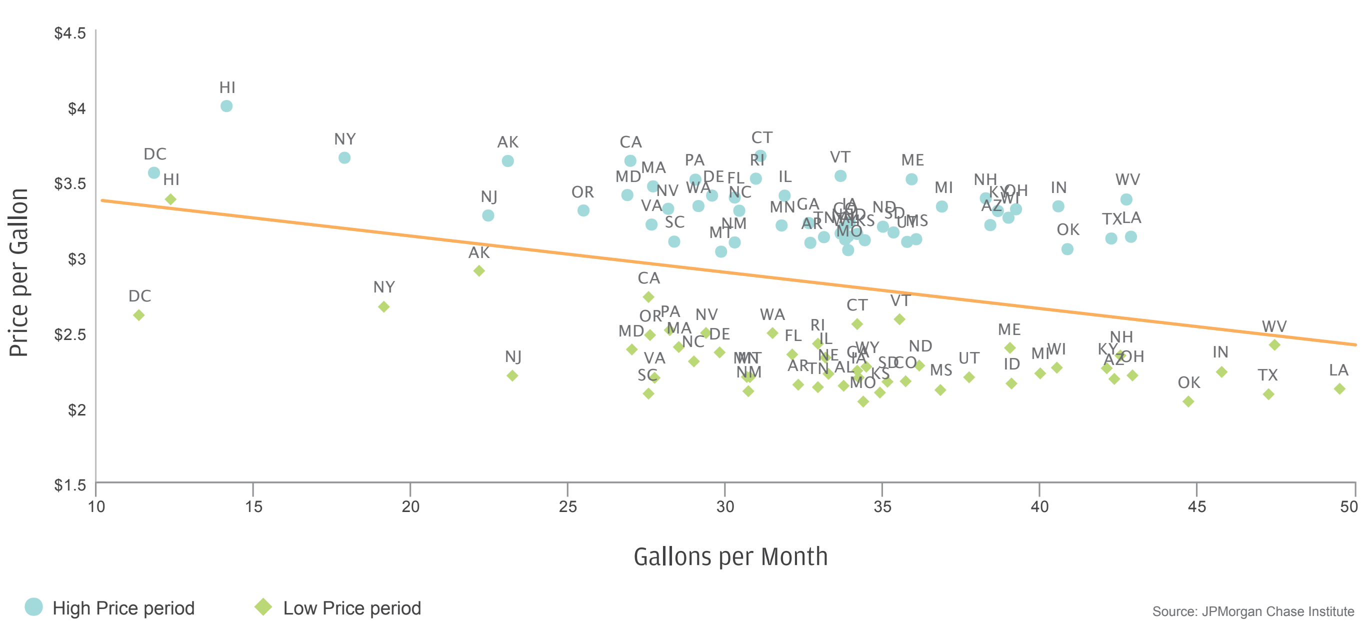
There was more than a two-fold spread in the percent change in price that states observed as gas prices fell nationally.

First, states varied significantly in terms of price levels. According to the EIA, in the High Price period, and in general, prices in California ($3.68 per gallon) and New York ($3.73 per gallon) were significantly higher than in Texas ($3.19 per gallon) and Colorado ($3.23). State differences in gas prices are largely due to differences in state gas tax rates, which vary from as high as 70 cents per gallon in Pennsylvania to as low as 31 cents per gallon in Alaska, as estimated by the American Petroleum Institute (2015).[12]

Second, there was more than a two-fold spread in the percent change in price that states observed as gas prices fell nationally between the High Price and Low Price periods. Six states experienced a price decline of more than one-third (Ohio, Michigan, Oklahoma, Texas, Missouri), whereas Hawaii saw only a 15% drop in gas prices. Moreover, states with higher relative gas prices saw smaller percent changes in gas prices (e.g., 25% price decline in California) during this period than states with low prices (e.g., 32% price decline in Texas). This is likely because differences in state gas prices are driven by gas taxes, which are largely implemented on a per gallon basis. This fixed-rate tax structure tends to dampen price fluctuations in percent terms in states with high gas taxes.

Third, in terms of quantity of gas, there is almost a four-fold spread in the amount of gas consumed per capita across states, with Louisiana, West Virginia and Texas at the top and New York, Hawaii and the District of Columbia at the bottom. As shown in Figure 10 below, in states that have higher prices, individuals consume less gas. This correlation reflects many factors other than price that influence gas consumption both across states and within each state over time.

Figure 10: Quantity of gas consumed and price of gas in the High Price period and the Low Price period, by state

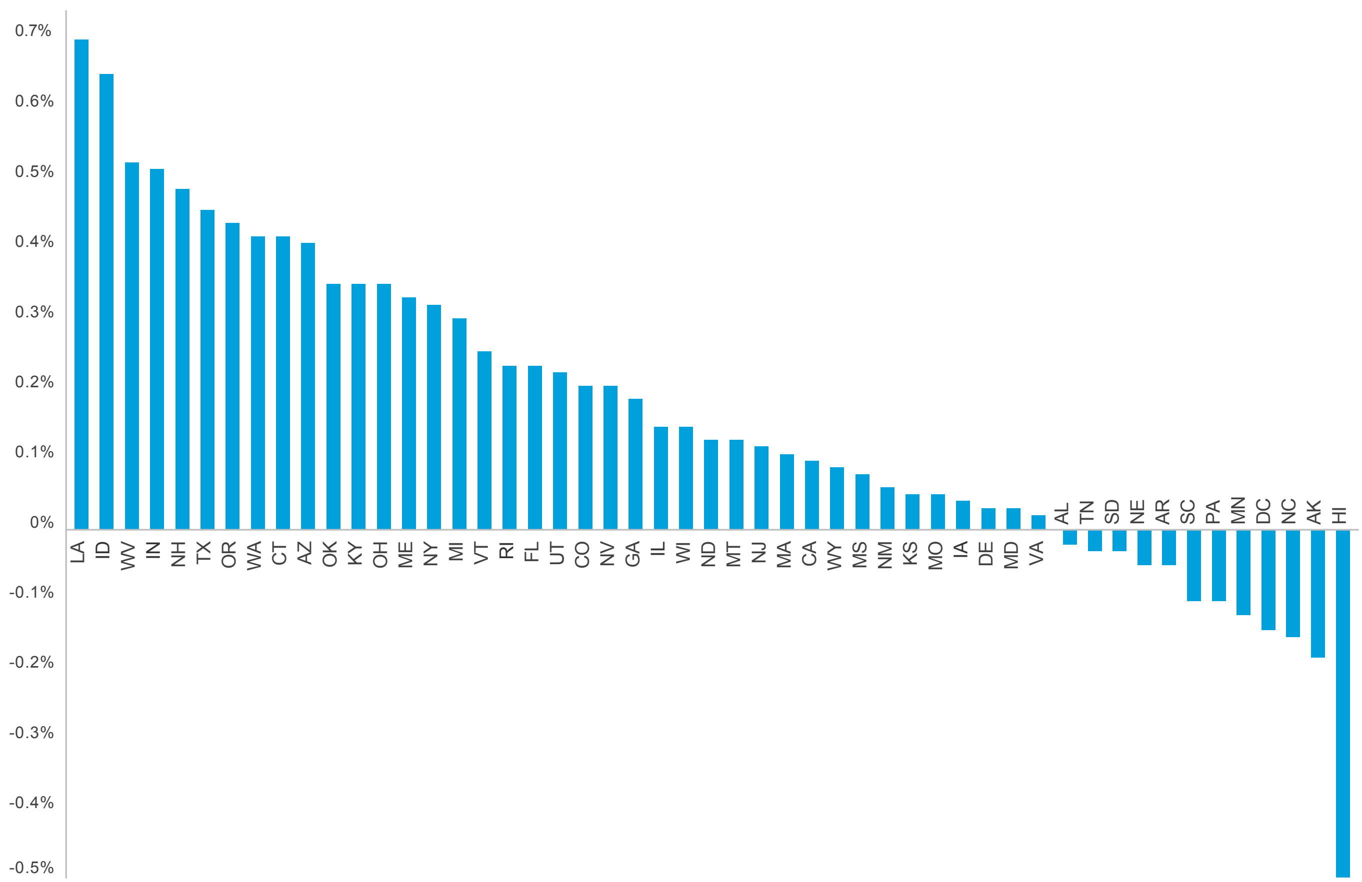


Source: JPMorgan Chase Institute

Finally, in Figure 11, we explore changes in the quantity of gas consumed in each state. On average, the quantity of gas consumed increased by almost 4% between the High Price period and the Low Price period. Fifteen states experienced more than a 0.3% increase in gas consumed for every 1% drop in gas prices between the High Price period and the Low Price period.[13] These are states where people already consume a lot of gas. Individuals in these states increased their gas consumption significantly when prices declined. Twenty-four states saw more moderate quantity drops as a fraction of price changes between the High Price and Low Price periods.[14] In the remaining 11 states and the District of Columbia we observed a decrease in gas consumption between the High Price period and the Low Price periods. These states had relatively low absolute levels of spending on gas and relatively low gas consumption when gas prices were higher. Individuals in these states further decreased their gas consumption even as gas prices declined. It is worth noting that all of these states are outside of Chase's branch footprint, which may influence their results.

The quantity of gas consumed increased by almost 4% between the High Price period and the Low Price period.

Figure 11: Percent change in quantity associated with a 1% decrease in price between the High Price period (Dec 2013–Feb 2014) and the Low Price period (Dec 2014–Feb 2015)



Source: JPMorgan Chase Institute

Bringing these four pieces together—gas price and consumption levels before the fall in gas prices and the resulting changes in each—allows us to break down the differences across states in the drop in gas spend displayed in Figure 9. Although price changes account for most of the change in spending, substantial quantity changes also occurred in some cases. Two segments of states emerge from these complex dynamics.

Figure 12: Higher-Impact states and Lower-Impact states during the gas price decline

	Initial Gas Prices	Initial Quantity of Gas Consumed	Drop in Gas Prices
Higher-Impact States Large Drop in Gas Spending **Midwest** IA IN KS MI MO OH SD **South** AL KY LA MS OK TN TX	Low	High	High
Lower-Impact States Small Drop in Gas Spending **East** CT DE DC FL MA MD NC NY PA **West** AK CA HI NV OR WA	High	Low	Low

Source: JPMorgan Chase Institute

First, there are the higher-impact states. These states consumed a lot of gas, tended to have lower prices, and saw larger drops in prices (e.g., more than a 30% drop). Most of these states saw relatively large drops in gas spending (Figure 9) despite that people in these states were more likely to use some of their gas savings to purchase more gas (i.e., they increased consumption more significantly for every 1% drop in price as shown in Figure 11). In Figure 10 these states moved both down and to the right significantly between the High Price and Low Price periods. Seventeen states fall into this segment, including much of the South (Alabama, Kentucky, Louisiana, Mississippi, Oklahoma, Tennessee, Texas) and the Midwest (Indiana, Iowa, Kansas, Michigan, Missouri, Ohio, South Dakota).

In the Midwest and South, higher-impact states, people saw the largest percentage declines in gas prices and gas spending as a fraction of income.

Then there are the lower-impact states. These states tend to be places where gas consumption is relatively low, prices are relatively high, and the change in price was relatively low (e.g., less than 30% drop). In Figure 10 these states moved down between the High Price and Low Price periods but not much to the right. Fifteen states fall into this segment, including much of the East (Connecticut, Delaware, District of Columbia, Florida, Maryland, Massachusetts, New York, North Carolina, Pennsylvania) and the West (Alaska, California, Hawaii, Nevada, Oregon, Washington). Perhaps not surprisingly they are precisely the areas that show the lowest change in gas spending, namely the East and West coasts of the country, as illustrated in Figure 9.

The states not included in either of these groups were in the middle in terms of drop in gas spending; this is because they were in the mid-range in terms of price, change in price, quantity and change in quantity. Bringing to light—for the first time to our knowledge—state-level data on all four of these dimensions illuminates how important these components are as state governments consider gas tax changes.

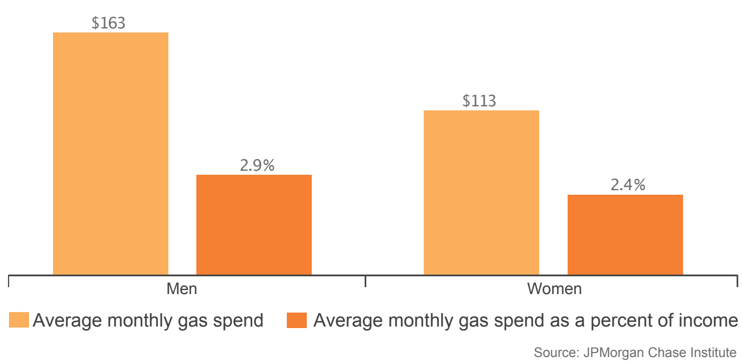
Finding Three

Savings at the gas pump represented more than 1% of monthly income for low-income individuals and disproportionately impacted younger Americans.

Men, individuals under 30 and those with low incomes spent the highest share of their income on gas

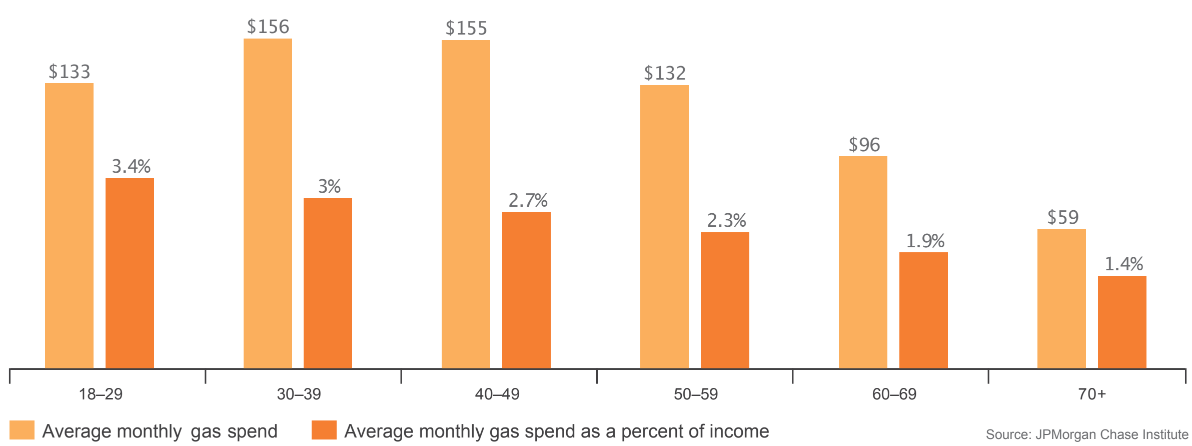
Next we examine demographic differences in gas spending behavior. We shift from our sample of 25 million consumers we used for our geographic analysis to a random sample of 1 million Chase customers whom we consider to be “core” and in whom we have greater confidence that we are seeing most of their spending activity (see the Data Asset and Methodology section for a description of the sampling criteria and characteristics of this sample). In absolute terms, men, individuals in their 30s and 40s, and high-income individuals spent the most on gas, but individuals under 30 and those with low incomes spent the largest share of their income on gas. The figures below show average gas spend in dollars and as a percent of income by gender, age and income. Men spent on average $163 per month on gas compared to $113 per month for women. As a fraction of their income, men spent slightly more (2.9%) on gas than women (2.4%).

Figure 13: Average monthly individual gas spend in the High Price period (Dec 2013–Feb 2014), by gender



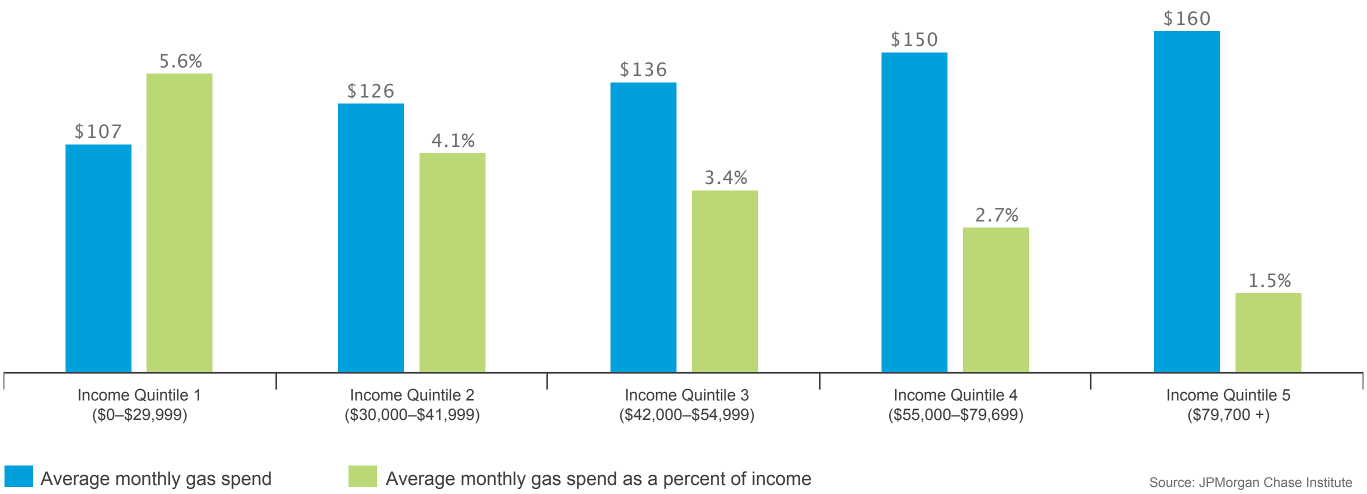
Gas spending is highest among individuals in their 30s and 40s, who spent around $155 per month on gas compared to less than $140 per month for all other age groups. In relative terms, individuals under 30 years old spent the most on gas as a fraction of their income (3.4%). Gas spending as a fraction of income declines steadily with age after age 30.

Figure 14: Average monthly individual gas spend in the High Price period (Dec 2013–Feb 2014), by age



Higher-income individuals spent more on gas in absolute terms, but low earners spent the most as a percent of their income. Those in the top income quintile (annual incomes greater than $79,900) spent $160 per month on gas, which translates to 1.5% of their monthly income. In contrast, those in the bottom income quintile spent only $107 per month on gas; yet, as a fraction of their monthly income they spend 5.6% on gas.

Figure 15: Average monthly individual gas spend in the High Price period (Dec 2013–Feb 2014), by income quintile

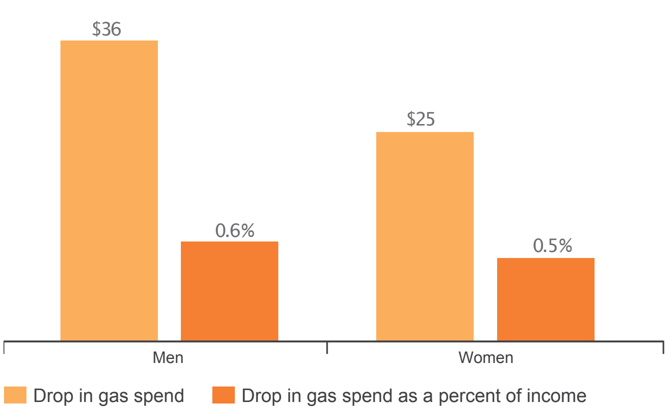


Source: JPMorgan Chase Institute

Gas savings represented more than 1% of monthly income for low-income individuals and disproportionately impacted younger Americans

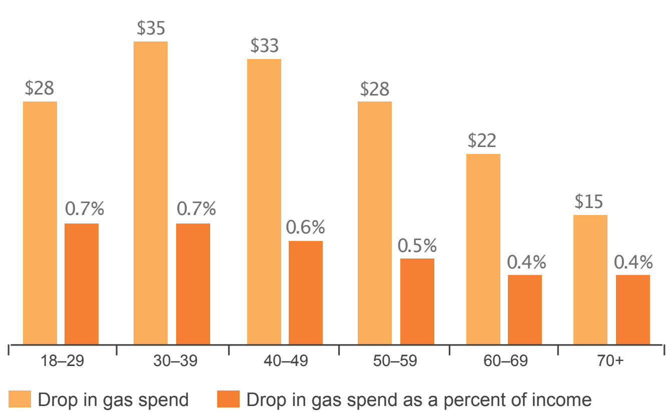
Next we explore who was most impacted by the price declines by examining how the change in gas spending varied across various demographic groups. Men, individuals in their 30s and high-income individuals saw the largest dollar value drop in gas spending. As a fraction of their income, men, individuals under 40, and low-income individuals saw the largest proportional increases in discretionary income as a result of spending less on gas.

Figure 16: Change in monthly gas spend between the High Price period (Dec 2013–Feb 2014) and the Low Price period (Dec 2014–Feb 2015), by gender



Source: JPMorgan Chase Institute

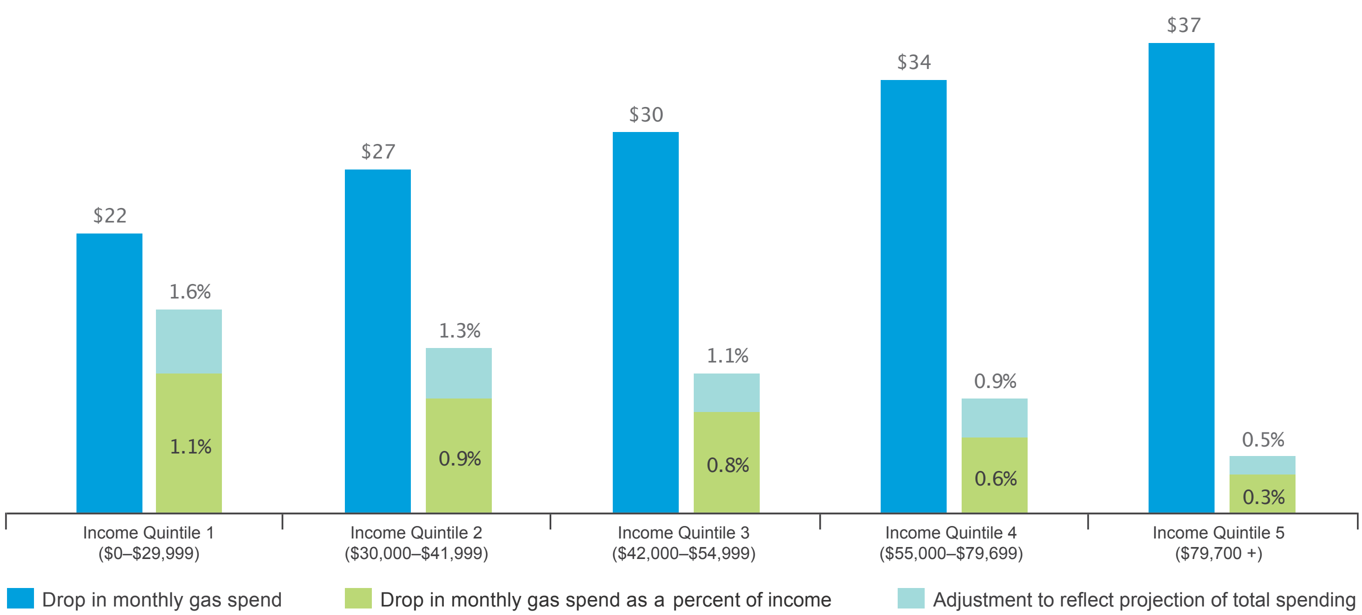
Figure 17: Change in monthly gas spend between the High Price period (Dec 2013–Feb 2014) and the Low Price period (Dec 2014–Feb 2015), by age



Source: JPMorgan Chase Institute

Low-income individuals saw the equivalent of a 1.1% increase in monthly income as a result of the decline in their gas spending, and middle-income individuals (quintiles 2 and 3) experienced the equivalent of a 0.9% and 0.8% increase in monthly income, respectively. Given that we estimate that we observe only 71% of gas spending, we adjust these figures to reflect a projection of total spending. Figure 18 below displays the drop in monthly gas spending as a percent of monthly income based on Chase credit and debit card spending as well as a projection of total spending. With this adjustment we see that low-income individuals experienced the equivalent of a 1.6% increase in income.[15] To put these numbers in perspective, between 2013 and 2014 bottom quintile individuals saw a 2.1% increase in income, according to the Current Population Survey. In other words, the increase in purchasing power lower-income people felt as a result of the decline in gas prices (1.6%) was equivalent to three-fourths of the increase in monthly income they experienced between 2013 and 2014 (2.1%).

Figure 18: Change in monthly gas spend between the High Price period (Dec 2013–Feb 2014) and the Low Price period (Dec 2014–Feb 2015), by income



Source: JPMorgan Chase Institute

In summary, we find that gas prices had disparate impacts across the United States. When gas prices decline, people living in the Midwest, men, those under age 40, and lower-income individuals experience the largest boost to their purchasing power.[16] We next explore whether and on what people spent their savings from lower gas prices.

Finding Four

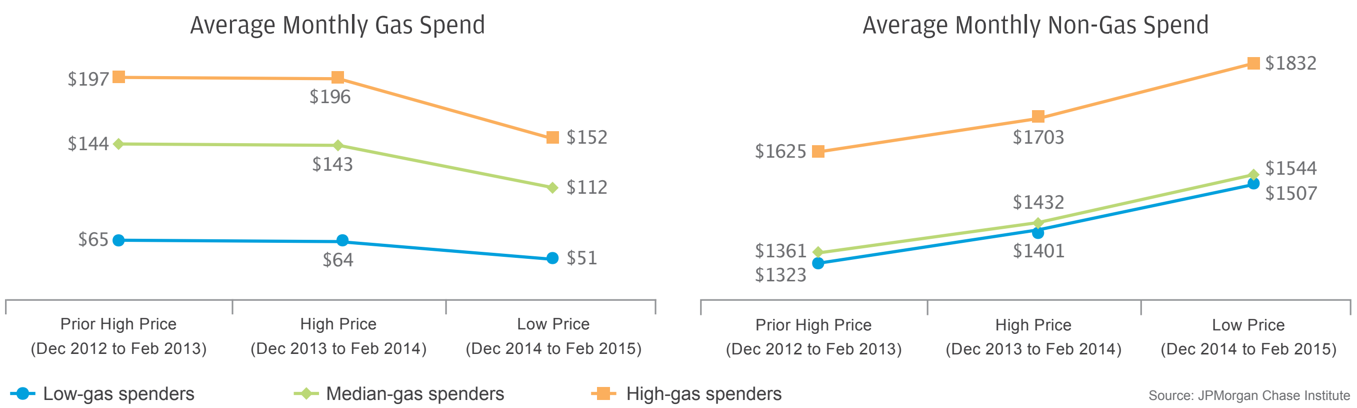
Individuals spent roughly 80% of their savings from lower gas prices.

The estimated marginal propensity to consume a dollar saved on gas is 73–89 cents or roughly 80%

Measuring the impact of the fall in gas prices on consumer spending is difficult to do with aggregate data because changes in non-energy consumption are potentially affected by many other economic factors. We isolate the causal impact of lower gas prices on non-gasoline spending by using anonymized individual-level spending data and comparing high-gas spenders to low-gas spenders.[17] We measure gas spending on debit and credit card transactions at gas stations and non-gas spending as all other transactions. Low-gas spenders are less impacted by gas price declines than high-gas spenders, yet they are affected similarly by other macroeconomic trends and market dynamics.[18] We validate low-gas spenders as a control group for high-gas spenders by comparing the spending behavior of these groups when gas prices were constant, between the High Price period (Dec 2013–Feb 2014; average price per gallon of $3.31) and one year prior (Dec 2012–Feb 2013; average price per gallon of $3.42), which we refer to as the Prior High Price period. In making these comparisons, it is worth noting that people who spend a lot on gas are not necessarily higher-income individuals. As shown in Figure 22 in the Data Asset and Methodology section, low-gas and high-gas spenders have comparable incomes.

In short, we expect many economic factors to affect everyone, but the decline in gas prices to disproportionately impact high-gas spenders.[19] This difference is evident in Figure 19 below, which displays average monthly gas and non-gas spending for high, median and low-gas spenders. When gas prices were steady, between the Prior High Price and High Price periods, there was little change in gas spending, while non-gas spending was increasing similarly for all groups.[20] In contrast, when gas prices dropped between the High Price and Low Price periods, gas spending dropped significantly more for high-gas spenders than for low-gas spenders ($45 compared to $13). Over the same time period, and facing similar economic conditions, high-gas spenders increased their non-gas spending by $23 more than low-gas spenders ($130 compared to $107). We use these figures to create our baseline estimate of the marginal propensity to consume (MPC)—for every additional dollar not spent on gas, individuals spent 73 cents ($23 of the $32 less in gas spending) on other things.

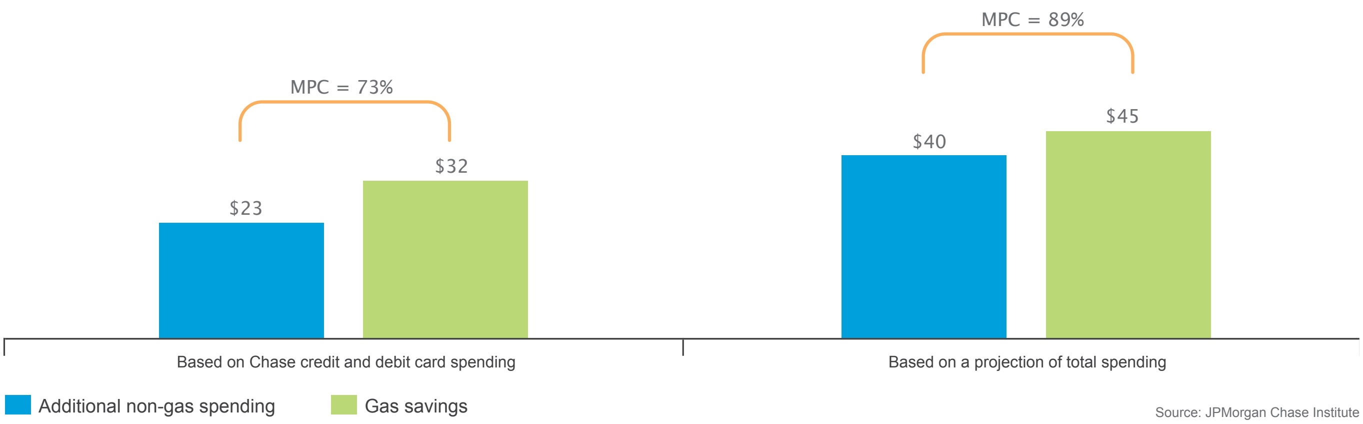
Figure 19: Levels of gas spending for low-gas spenders and high-gas spenders



We perform a number of robustness checks to this baseline MPC estimate of 73%, which we describe in the Data Asset and Methodology section. The bottom line, presented in Figure 24, is that our baseline MPC estimate of 73% is robust to a range of specifications, ranging from 60% to 74%. It increases to 89% when we adjust for the share of spending we observe on Chase debit and credit cards (estimated at 71% for gas spending and 58% for non-gas spending).[21] Even so, our results may underestimate the full extent to which people are spending their gains from lower gas prices, since we only consider spending categories that would ever appear on credit or debit cards. The range of these MPC estimates are higher than what has been implied by existing recent evidence, widely held beliefs and self-perceptions that people were not spending most of their gains at the pump. They are more in line, though, with historical evidence of a marginal propensity to consume that can even exceed 100%.[22]

The evidence suggests that gas price fluctuations indeed have a significant effect on consumer spending. Although the MPC that we measure is the individual consumption response to gas price declines, it may reflect not only individual consumer spending decisions but also short-term general equilibrium impacts within the local economy. For example, lower gas prices not only impact disposable income directly (as estimated in this report), but they also boost consumer confidence, decrease the operating costs of vehicles, drive down costs for businesses, and generate increased demand for, and therefore the production of, gasoline. To the extent that high-gas and low-gas spenders are concentrated geographically, and that these additional effects reverberate within these geographies in the short term, our estimate will include them as well.[23]

Figure 20: Marginal propensity to consume (MPC) from a $1 less spent on gasoline

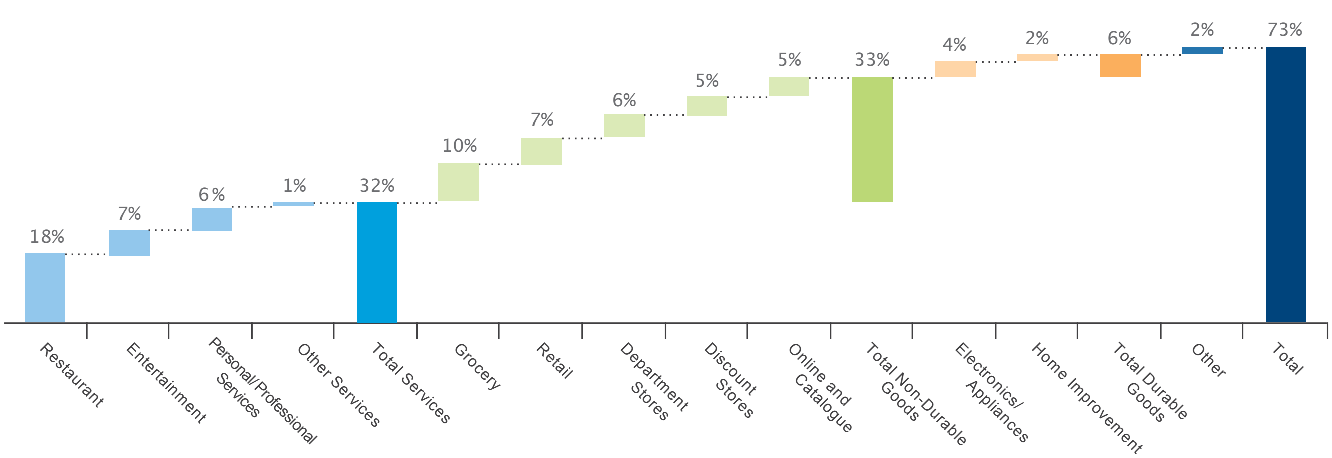


Source: JPMorgan Chase Institute

People spent almost 20% of their gas savings on restaurants alone

As a final step, we provide a more in-depth look into how people spent their marginal dollars gained from gas price declines. In Figure 21, we show the breakdown of which categories people spent their gas savings on. We see that people spent 18% of their gas savings on restaurants alone and a total of 32% on services. The largest share of savings on gas (33%) was spent on non-durable goods, with 10% spent on groceries alone. An additional 6% was spent on durable goods, and 2% on other categories (e.g., charitable donations). The categories that saw the largest growth in percentage terms were department stores (8% increase), entertainment (7% increase) and electronics and appliances (6% increase). It is important to note that this distribution of observed spending by category may be influenced by the tendency to use credit and debit cards in each category.[25]

Figure 21: Percent of savings from lower gas prices spent on non-gas categories, by spending category



Source: JPMorgan Chase Institute

Implications and Conclusions

We believe this study contributes to our understanding of how the recent gas price declines are impacting the U.S. consumer. First, contrary to general perception, people appear to be spending rather than saving their gains at the pump. Our estimates of the marginal propensity to consume are more in line with historical estimates of the impacts of gas price fluctuations on the economy, which show a larger effect of gas price fluctuations. We show a marginal propensity to consume in the range of 73%–89% when projecting total spending on debit and credit card spending categories. This estimate implies that consumers are mostly redistributing their gains at the pump to other spending categories. These estimates run contrary to correlational evidence of the impacts of gas price declines as well as self-reported perceptions of how consumers believe they are responding. Consumers report that they are using their gains at the pump to pay down debts and save. Our data show they are spending most of them.

Given that gas spending represents less than 5% of consumer spending, these impacts are small in absolute dollar terms and easily overshadowed by other economic forces. Nonetheless this boost to other categories of consumer spending could be here to stay if gas prices remain low as predicted. On the other hand, a substantial increase in gas prices might proportionately dampen consumer spend in these categories, if the response to gas price increases is symmetrical with the response to gas price decreases.

In addition, we show how gas price decreases have disparate impacts across the country: people in the Midwest and South, and the young and poor, feel the largest gains relative to their current income. For low-income earners, the recent gas price decline was equivalent to more than 1% of their income. This highlights the fact that gas price fluctuations contribute to spending volatility. Reduced reliance on gas, for example through electrification of the transportation sector, could reduce volatility particularly for low-income earners. In addition, innovative financial services could assist consumers in hedging gas price volatility. For example, to assist consumers in saving, credit cards (e.g., especially gas rewards cards) could develop a savings feature that activates when gas prices drop. Reducing volatility is an important goal, given that, as shown in our previous report *Weathering Volatility*, individuals across the income spectrum experience significant income and spending volatility and lack a sufficient financial buffer to withstand this volatility (Farrell and Greig, 2015).

Finally, the distributional impacts of gas price changes are important considerations for gas tax policy at the national and state level. Across the board, gas taxes are regressive, but gas taxes based on quantity consumed (rather than price) mitigate gas price fluctuations in percentage terms. Efforts to increase gas taxes should consider ways to make these taxes more progressive in order to mitigate the impact on those with lower incomes. We show that states differ dramatically in terms of price levels, price changes, quantity levels and quantity changes. Taken together, states in the Midwest and South were far more impacted by gas price declines than states in the East and West coasts. In the Midwest and South, higher-impact states, people saw the largest percentage declines in gas prices and gas spending as a fraction of income, despite the fact that residents of these states increased their driving the most for each 1% decline in price. In the East and West, lower-impact states, people saw smaller drops in gas prices and gas spending as a fraction of income. Our results imply that increasing the gas tax in higher-impact states might increase tax revenue and also lead people to drive less. Conversely, in the lower-impact states, where people tend to pay high prices and consume less gas to begin with, an increase in the gas tax might yield increased tax revenue without curbing gas consumption. These state-level differences provide a more granular understanding of how gas price fluctuations impact regional economies and should inform good decisions about optimal gas tax rates and structures.

When we embarked on this research project, the prevailing wisdom was that consumers were using their gas savings to repair their balance sheets, perhaps because they viewed the price declines as temporary or were suffering from a "debt overhang." We present evidence that recent gas price declines resulted in significantly more spending than previously understood, and that the gains in discretionary spending disproportionately accrue to low-income individuals, to young people, and to states where people spend a lot on gas. This is good news for the U.S. consumer as we anticipate sustained low gas prices through the rest of 2015.

The JPMorgan Chase Institute Data Asset and Methodology

In this report, the JPMorgan Chase Institute seeks to inform the public debate on the impact of the recent gas price declines on consumer spending. To develop insights into these topics, we adapted the Bank's internal consumer data on 57 million anonymized U.S. debit and credit cardholders into a groundbreaking data asset. As the first financial institution to channel this wealth of information for the benefit of the public good, JPMorgan Chase & Co. put strong guardrails and strict privacy protocols in place to protect personal information throughout the creation and analysis of this data asset. A description of these protocols are available on our website.

Data Privacy

The JPMorgan Chase Institute has adopted rigorous security protocols and checks and balances to ensure all customer data are kept confidential and secure. Our strict protocols are informed by statistical standards employed by government agencies and our work with technology, data privacy and security experts who are helping us maintain industry-leading standards.

There are several key steps the Institute takes to ensure customer data are safe, secure and anonymous:

- Before the Institute receives the data, all unique identifiable information—including names, account numbers, addresses, dates of birth and Social Security numbers—is removed.
- The Institute has put in place privacy protocols for its researchers, including requiring them to undergo rigorous background checks and enter into strict confidentiality agreements. Researchers are contractually obligated to use the data solely for approved research, and are contractually obligated not to re-identify any individual represented in the data.
- The Institute does not allow the publication of any information about an individual consumer or business. Any data point included in any publication based on the Institute's data may only reflect aggregate information.
- The data are stored on a secure server and can be accessed only under strict security procedures. The data cannot be exported outside of JPMorgan Chase's systems. The data are stored on systems that prevent them from being exported to other drives or sent to outside email addresses. These systems comply with all JPMorgan Chase Information Technology Risk Management requirements for the monitoring and security of data.

The Institute provides valuable insights to policymakers, businesses and nonprofit leaders. But these insights cannot come at the expense of consumer privacy. We take precautions to ensure the confidence and security of our account holders' private information.

Constructing our Sample

For this report we rely on JPMorgan Chase data on consumer clients who are primary account holders. To avoid double counting of financial activity, all joint accounts are captured under one individual, the primary account holder. From a universe of over 57 million anonymized debit or credit card account holders nationwide, we created a sample of 25.6 million individuals who we believe to be regular users of a Chase credit or debit card. We selected individuals who have an average of five transactions a month on either their credit or debit card. We use this vast population to conduct all of our geographic analyses (Finding 2), as it provides broad coverage of the nation. Our maps report statistics for any county in which we have a minimum of 50 customers who have on average five transactions a month—roughly 95% of counties.

As shown in Figure 22, this population of 25.6 million is different from the nation in important ways. First, our sample is skewed slightly in favor of younger individuals: it slightly over represents individuals aged 30-49 and underrepresents individuals over age 70. Second, the JPMC Institute sample includes a high proportion of men. This bias may reflect a tendency for men to be listed as primary account holders on joint accounts rather than an underlying bias in the Chase population in favor of men. Third, our sample is biased geographically by Chase's footprint, which gives us broad coverage of the four Census regions, but with a slight bias in favor of the West, when compared to the nation. Finally, our sample is skewed in favor of higher-income individuals for a number of reasons. In our data asset, we observe only those individuals who have a relationship with Chase. Roughly 8% of Americans do not bank with a U.S. financial institution and tend to be disproportionally lower-income and non-Asian minorities (FDIC 2014).

Figure 22: Demographic characteristics of JPMorgan Chase Institute sample versus the U.S. population

		JPMC Institute Sample					
	U.S. Population[1]	25.6 million sample[4]			1 million sample[5]		
		All	Debit Card Holders	Credit Card Holders	All	Low-Gas Spenders	High-Gas Spenders
18-29	22%	20%	29%	13%	21%	26%	18%
30-39	17%	19%	21%	17%	24%	26%	24%
40-49	17%	19%	20%	19%	22%	20%	24%
50-59	18%	19%	17%	21%	18%	15%	19%
60-69	14%	14%	9%	18%	10%	9%	11%
70+	12%	9%	4%	13%	5%	4%	5%
Men	49%	53%	53%	N/A[6]	53%	49%	56%
Women	51%	47%	47%	N/A[6]	47%	51%	44%
Northeast	18%	19%	19%	22%	19%	73%	3%
Midwest	21%	21%	19%	22%	19%	10%	27%
South	37%	30%	28%	30%	28%	4%	48%
West	23%	30%	34%	26%	34%	12%	22%
Monthly Income	$3,626[2]	$6,020	$4,811	$7,286	$5,085	$5,318	$5,283
Monthly Gas Spend	$206[3]	$105	$124	$74	$146	$65	$210
Monthly Non-Gas Spend	$2,525[3]	$1,295	$1,105	$1,340	$1,524	$1,479	$1,768

Source: JPMorgan Chase Institute

1 Unless otherwise noted, national estimates come from the Census Bureau's American Community Survey 2013 1 Year Estimates.

2 Estimates are from the 2014 Current Population Survey and represent person income estimates.

3 Estimates come from the 2014 Consumer Expenditure Survey. Non-gas spend excludes categories of spending that are unlikely to be conducted using a debit or credit card, specifically: auto purchase, auto finance, gas, shelter, and pension.

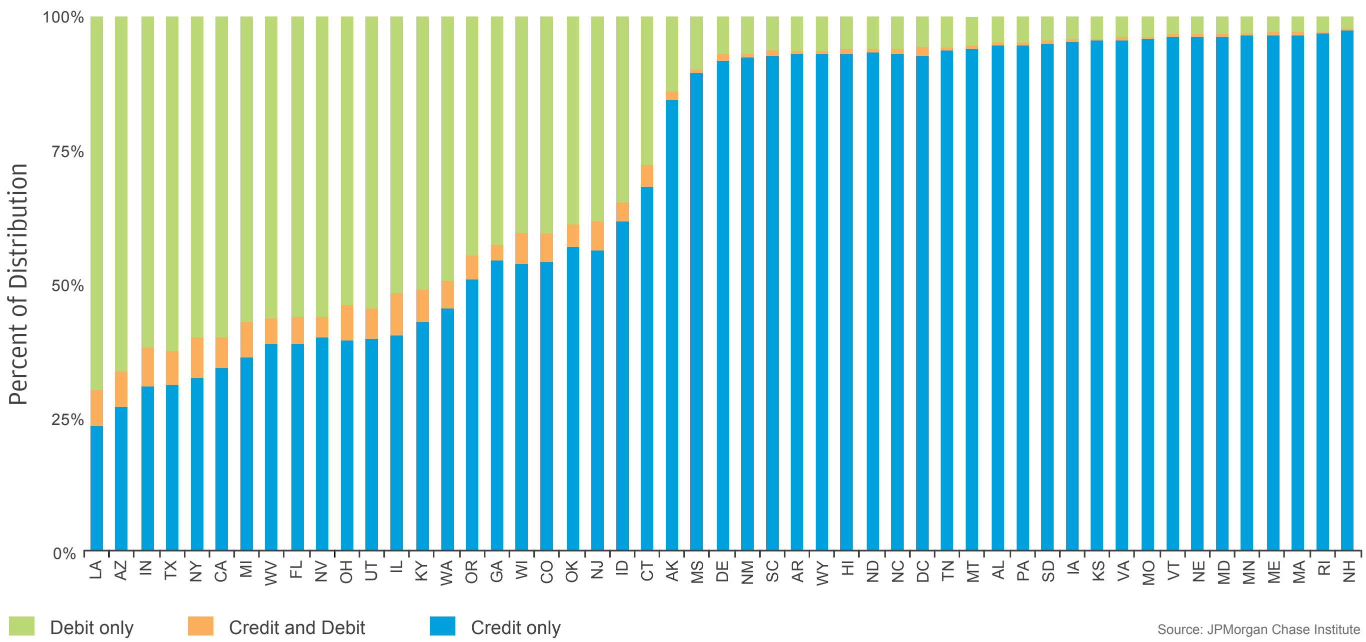
4 The 25.6 million sample includes individuals who have either a credit or debit card and an average of five transactions a month on either one. This sample is used for our geographic analyses in Finding 2.

5 The 1 million sample includes checking account holders with a minimum of five outflows per month who do not have a gas station specific Chase credit card, and who live in a zip code with at least 140 other individuals in our sample.

6 Gender information is not available for credit card holders.

In addition to the differences between our population and the nation in aggregate, there may be additional biases at the state level. As shown in Figure 23, the distribution of credit and debit card users varies dramatically across states. This is due to the fact that Chase's credit card presence spans the nation, whereas checking accounts can only be opened within the 23 states in which Chase has physical branches. Given that debit cardholders tend to be significantly younger and have lower incomes than credit card users, the distributions below may influence the levels of gas and non-gas spending observed by state. Figure 22 presents the demographic characteristics separately for debit card versus credit cardholders in our 25.6 million sample used for our geographic analyses.

Figure 23: Distribution of credit and debit cardholders in 25.6 million sample, by state



Source: JPMorgan Chase Institute

For all other analyses in this report (Findings 1, 3 and 4), we construct a one million person sample that gives us greater confidence that we are seeing most of an individual's spending activity. We apply a more stringent criteria to identify individuals we believe are "core" Chase customers and conduct most of their gas and non-gas spending behavior using a Chase debit or credit card. Specifically, we take a random sample of 1 million debit cardholders who meet the following additional sampling criteria:

1. They have a checking account and at least 5 outflow transactions from their checking account per month.
2. They do not hold a gas station specific card.
3. They live in a zip code with at least 140 other individuals in our sample.

These additional criteria give us confidence that we are focusing on core Chase clients as we assess the impact of low gas prices on spending behavior. These criteria constrain our sample to the 23-state Chase branch footprint within the nation. As shown in Figure 22, the 1 million sample is even more skewed towards younger individuals than the 25 million geographic sample, but it is more representative of the nation in terms of income.

Drawing from a universe of over 57 million anonymized customers, we sampled 25 million regular Chase credit and debit card users to shed new light on the effects of gas price decreases on consumer spending.

Measuring Spending

We measure spending behavior using debit and credit card transactions, which we refer to as card spending. We focus exclusively on card spending because we are able to clearly distinguish between gas and non-gas spending. Specifically, we analyze merchant information of these transactions and classify all card spending at gas stations, including attached convenience stores, as "gas spending" and all other card spending (i.e., not at gas stations) as "non-gas spending."[26] Card spending offers clean, albeit incomplete, measures of gas and non-gas spending. Card spending provides a relatively good window into spending on goods and services but less visibility into spending categories where individuals more frequently use cash, checks and electronic transfers, such as rent payments, utility bills and vehicle purchases.[27]

Estimating the marginal propensity to consume

People spent almost 20% of their gas savings on restaurants alone.

We use a "difference in difference" approach to isolate the impact of low gas prices on consumer spending from other economic and market conditions and trends over this timeframe. Specifically, we compare the difference between high-gas and low-gas spenders in their difference in non-gas spending between the High Price period (Dec 2013–Feb 2014) and the Low Price period (Dec 2014–Jan 2015). In this research design, our low-gas spenders serve as a control group for how high-gas spenders would have behaved had gas prices not dropped. We believe low-gas spenders are a valid control group because, as indicated in Figure 19, high-gas spenders and low-gas spenders showed very similar trends between the Prior High Price period (Dec 2012–Feb 2013) and High Price period (Dec 2013–Feb 2014), when gas prices were high and relatively constant. The spending patterns only diverge when we move from the the High Price period (Dec 2013–Feb 2014) to the Low Price period (Dec 2014–Jan 2015).

For this analysis, we assign individuals as either low-gas spenders (bottom quintiles of gas spend) or high-gas spenders (top quintile of gas spend), but categorize each individual based on the mean gas spend in their zip code, excluding their own gas spend. We assign individuals to gas spend quintiles using this zip code level "leave-out mean" in order to prevent our results from being biased by mean reversion in individual gas spending over time.[28] Using the leave-out mean to assign people to quintiles of gas spend does not significantly change the demographic or economic characteristics of the individuals categorized as low-gas versus high-gas spenders.

Figure 24 displays the 95% confidence interval for our "baseline" estimate of the marginal propensity as explained above, as well as a variety of robustness checks and adjustments to this estimate. We estimate the 95% confidence interval for the marginal propensity to consume estimate through an instrumental variable regression approach, in which we use whether a person is a high-gas versus low-gas spender (assigned based on the leave-out mean gas spend in individual i's zip code) as an instrument for the year-over-year change in gas spend (Equation 1 below). We then regress year-over-year change in non-gas spend on the predicted year-over-year change in gas spend (Equation 2).

$$(1)\ \Delta Gas\ Spend_i = a_i + \beta_1 * I(High\ Gas\ Spender)_i + \varepsilon_i$$

$$(2)\ \Delta Non\ Gas\ Spend_i = a_i + \beta_2 * \Delta\widehat{Gas\ Spend}_i + \varepsilon_i$$

We define the dependent variables in these equations as:

$$(3)\ \Delta Gas\ Spend_i = GasSpend_i^{LowPrice} - GasSpend_i^{HighPrice2}$$

$$(4)\ \Delta Non\ Gas\ Spend_i = NonGasSpend_i^{LowPrice} - NonGasSpend_i^{HighPrice2}.$$

The coefficient of interest is β_2 in equation 2 above, which represents the marginal propensity to consume—the ratio of the difference in the change in non-gas spending for high-gas spenders versus low-gas spenders. This IV estimate is equivalent to simply dividing the difference-in-difference estimate of the impact on non-gas spending by the estimate of the impact on gas spending. We estimate the 95% confidence interval by multiplying the standard error of β_2 by ±1.96.

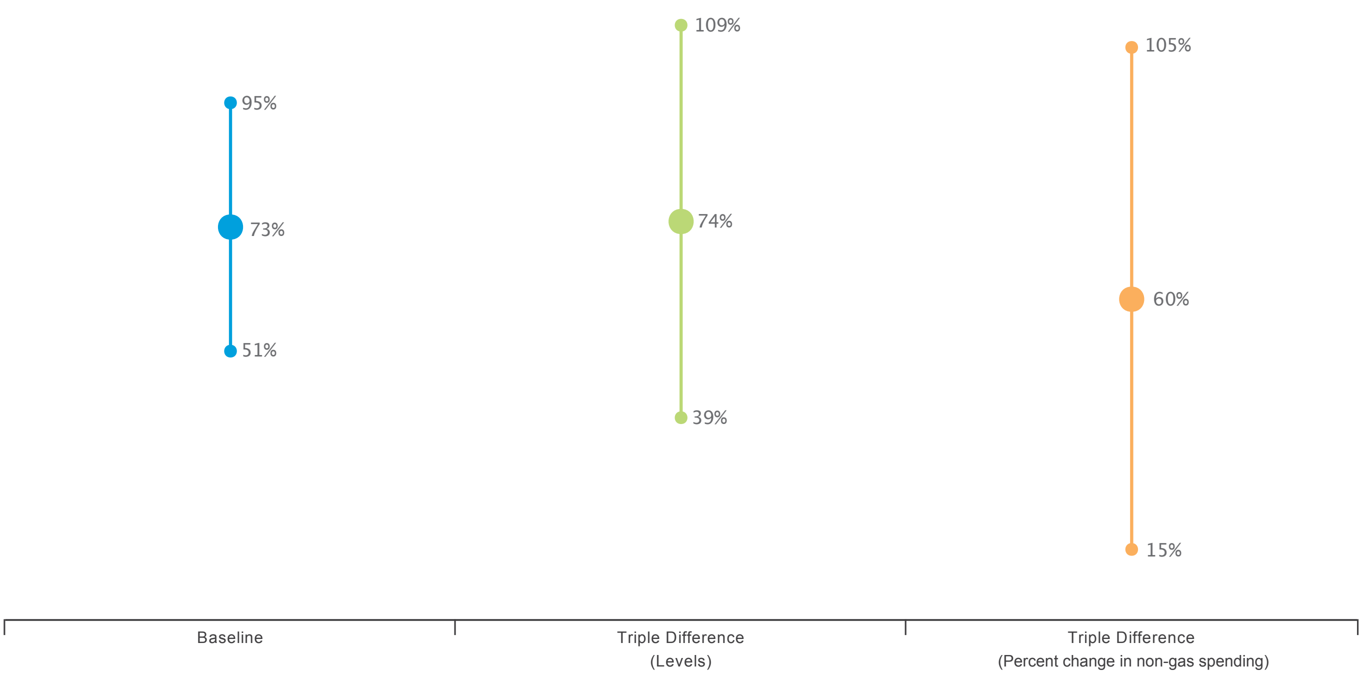
We conduct two different robustness checks to adjust our baseline estimates to account for the trends in gas and non-gas spending that were occurring even while gas prices were stable. When describing these adjustments, we refer to changes in the “pre” period as changes between the Prior High Price period (Dec 2012–Feb 2013) and the High Price period (Dec 2013–Feb 2014), and we refer to changes in the “post” period as changes between the High Price period and the Low Price period (Dec 2014–Feb 2015). The point of the robustness checks is to account for underlying changes between the high-gas and low-gas spending groups during the pre period.

First, our “Triple Difference—Levels” estimate adjusts for these trends in absolute terms. This is done by subtracting the dollar change in spending in the pre-period from our calculation of the dollar change in spending during the post period. This estimate effectively removes the pre-trends in dollar terms, and is valid assuming that these pre-trends would have continued similarly for both groups in the absence of gas price changes.

Second, our “Triple Difference—Percent Change in Non-Gas Spending”, estimate recovers the MPC by first estimating the effect on non-gas spending in percent terms.[29] This is done by using the “Triple Difference—Levels” method described above, except with each difference calculated in percent rather than dollar terms. This gives an estimate of the impact on non-gas spending in percent terms. To recover a dollar estimate, we then multiply this by the level of gas spending in the High Price period (Dec 2013–Feb 2014). Finally, to calculate the MPC, we divide this by the difference-in-difference estimate for gas spending calculated in dollar terms (since the percent impact is similar between both groups due to the price change). This estimate removes the pre-trend for non-gas spending in percent terms more directly than the estimate above, but uses the simple difference-in-difference estimate for gas spending.

We believe both of these adjustment approaches are instructive given that high-gas spenders spend 21% more than low-gas spenders on non-gas categories, and that the differences in pre-trends are substantial in percentage terms and still there, though minor, in dollar terms. Below, we show each of the three estimates of the marginal propensity to consume. Unadjusted for the share of total spend on credit and debit cards, the point estimate for the marginal propensity to consume ranges from 60% to 74% with a 95% confidence interval across the three estimates ranging from 15% to 109%.

Figure 24: Estimated marginal propensity to consume and 95% confidence intervals



Source: JPMorgan Chase Institute

In order to more closely represent the impact of low gas prices on the purchase of goods and services generally in the economy, we scale our three estimates to account for the fact that people pay for a higher share of their total gas spending using a debit or credit card (versus checks, electronic payments or cash) relative to non-gas card spending categories. Comparing observed per person spend on Chase cards relative to total expenditures reported in the Consumer Expenditure Survey, we estimate that roughly 71% of gas spending occurs on debit and credit cards and only 58% of non-gas spending occurs on debit and credit cards.[30] This adjustment requires that we multiply all of our point estimates in Figure 24 by 1.2 (the ratio of 71% and 58%), which shifts the MPC range from 60%–74% up to 74%–91%.

For every dollar less spent at the gas pump, individuals spent roughly 80 cents on other things.

Finally, we explore possible heterogeneity in our results by calculating the MPC separately by income quintile and by Census region. Although we find that low-gas spenders and high-gas spenders have similar income levels (see Figure 22), we might expect to see a higher marginal propensity to consume among lower-income individuals. The implied MPCs from the change in mean gas and non-gas spending for the bottom through the top income quintiles were 90%, 27%, 74%, 126%, and 72%, respectively. After adjusting for trends in non-gas spending between the Prior High Price period and the High Price period, each of the MPCs falls between 98% and 114%, with the exception of the second income quintile in which the MPC was negative. Thus we do not find evidence for the hypothesis that lower-income individuals have a higher propensity to consume their savings at the pump. Moreover, due to the instability of the MPCs after adjusting for pre-trends, we do not find these findings conclusive.

Similar problems arise when attempting to estimate MPCs for each Census region. The implied MPC was 56% for the Midwest, 56% for the Northeast, 199% for the South, and 72% for the West. With the exception of the South, each region exhibited stable pre-trends in gas spending between the Prior High Price and High Price periods. In terms of non-gas spending, only the Midwest exhibited pre-trends that would have allowed for a valid difference-in-difference framework. While we do estimate marginal propensities to consume after adjusting for these pre-trends, there is considerable noise associated with these estimates that makes them inconclusive.

In summary, we find robust estimates of a marginal propensity to consume ranging from 60% to 74% for the nation as a whole, which after accounting for the full range of spending beyond credit and debit card transactions scale up to a range of 74% to 91%.

Future enhancements of JPMorgan Chase Institute Data Assets

Our new and evolving consumer finance data asset provides fresh insights into the impacts of the recent declines in gas prices on consumer behavior. The JPMorgan Chase Institute will continue to build and refine this data asset to address an even broader array of important economic and policy questions pertaining to consumers and households. Ultimately, our ability to understand where consumers spend their money and how this varies month to month is an important cornerstone of our data asset. Other planned expansions to the data asset include a more complete view of consumer assets and liabilities to develop a perspective on household balance sheets. Finally, while still preserving the anonymity of our data, we plan to add third-party data on demographics to develop a granular perspective on consumer finance issues by important segments of the population and household characteristics.

In addition to our consumer data asset, the future research agenda of the JPMorgan Chase Institute extends across the portfolio of JPMorgan Chase's lines of business and vast geographic reach. Future data assets and analytics of the JPMorgan Chase Institute will focus on businesses large and small, the global flows of funds, and other critical economic topics. These data, combined with expert insights, are unique assets the JPMorgan Chase Institute will use to provide a comprehensive perspective on the complex inner workings of the global economy and help policymakers, businesses and nonprofit leaders make smarter decisions to advance global prosperity.

References

Bennet, B., Conover, D., O'Brien, S., & Advincula, R. (2014). Cash continues to play a key role in consumer spending: Evidence from the Diary of Consumer Payment Choice. Retrieved from http://www.frbsf.org/cash/publications/fed-notes/2014/april/cash-consumer-spending-payment-diary.

Coglianese, J., Davis, L. W., Kilian, L., & Stock, J. H. (2015). Anticipation, tax avoidance, and the price elasticity of gasoline demand (No. w20980). National Bureau of Economic Research.

Council of Economic Advisors (2015). Explaining the U.S. petroleum surprise. Council of Economic Advisors.

Edelstein, P., & Kilian, L. (2007). Retail energy prices and consumer expenditures. CEPR Discussion Papers 6255, C.E.P.R. Discussion Papers.

Edelstein, P., & Kilian, L. (2009). How sensitive are consumer expenditures to retail energy prices? *Journal of Monetary Economics*, 56(6), 766-779.

Farrell, D., & Greig, F. (2015). Weathering Volatility: Big Data on the Financial Ups and Downs of U.S. Individuals. JPMorgan Chase Institute.

Federal Deposit Insurance Company (2014). 2013 FDIC National Survey of Unbanked and Underbanked Households. Federal Deposit Insurance Company.

Federal Reserve Board of Governors (2015). Transcript of Chair Yellen's Press Conference: June 17, 2015. Retrieved from http://www.federalreserve.gov/mediacenter/files/FOMCpresconf20150617.pdf.

Furman, J. (2015). Advance Estimate of GDP for the First Quarter of 2015. Retrieved from https://www.whitehouse.gov/blog/2015/04/29/advance-estimate-gdp-first-quarter-2015. Council of Economic Advisors.

Gagnon, J., Raskin, M., Remache, J., & Sack, B. P. (2010). Large-scale asset purchases by the Federal Reserve: did they work? Federal Reserve Board of New York Staff Report, (441).

Gicheva, D., Hastings, J., & Villas-Boas, S. (2010). Investigating Income Effects in Scanner Data: Do Gasoline Prices Affect Grocery Purchases? *The American Economic Review*, 480-484.

Hamilton, J. D. (2009). Causes and Consequences of the Oil Shock of 2007-08 (No. w15002). National Bureau of Economic Research.

Hill, C. (2015). Used car prices hit a record high—but that's good news for some buyers. Retrieved from http://www.marketwatch.com/story/used-car-prices-hit-a-record-high-but-thats-good-news-for-some-buyers-2015-08-21.

Kelley Blue Book (2015). New-Car Transaction Prices Soar In January 2015, Up More Than 5 Percent, According To Kelley Blue Book. Retrieved from http://mediaroom.kbb.com/new-car-transaction-prices-soar-january-2015.

Institute on Taxation and Economic Policy (2011). Building a Better Gas Tax: How to Fix One of State Government's Least Sustainable Revenue Sources. Retrieved from http://www.itepnet.org/bettergastax/bettergastax.pdf.

Paszkiewicz, L. (2003). The Cost and Demographics of Vehicle Acquisition. *Consumer Expenditure Survey Anthology* (61).

Patrabansh, S., Doerner, W. M., & Asin, S. (2014). The Effects of Monetary Policy on Mortgage Rates. Federal Housing Finance Agency Working Paper (14-2).

Perks, R. & Raborn, C. (2013) Driving Commuter Choice in America Expanding Transportation Choices Can Reduce Congestion, Save Money and Cut Pollution. *National Resource Defense Council Issue Paper* (13-06-A).

Santos, A., McGuckin, N., Nakamoto, H. Y., Gray, D., & Liss, S. (2011). Summary of travel trends: 2009 national household travel survey. U.S. Department of Transportation, Federal Highway Administration (No. FHWA-PL-ll-022).

Swift, A. (2015). *Most in U.S. Say Low Gas Prices Make Difference in Finances*. Gallup. Retrieved from http://www.gallup.com/poll/183596/say-low-gas-prices-difference-finances.aspx.

U.S. Energy Information Administration (2015a). Gasoline prices tend to have little effect on demand for car travel. Retrieved from http://www.eia.gov/todayinenergy/detail.cfm?id=19191.

U.S. Energy Information Administration (2015b). Short-Term Energy Outlook September 2015. Retrieved from http://www.eia.gov/forecasts/steo/report/prices.cfm.

U.S. Energy Information Administration (2015c). Unplanned refinery outage leads to higher Midwest gasoline prices. Retrieved from http://www.eia.gov/todayinenergy/detail.cfm?id=22552.

U.S. Energy Information Administration (2015d). U.S. household gasoline expenditures expected to fall in 2015. Retrieved from http://www.eia.gov/todayinenergy/detail.cfm?id=20752.

Vock, D.D. (2015). Raising Gas Taxes Gets Bipartisan Boost from Governors. Retrieved from http://www.governing.com/topics/transportation-infrastructure/gov-low-gas-prices-deteriorating-roads-fuel-push-for-gas-tax-hikes.html.

Endnotes

1 See U.S. Energy Information Administration (2015b and 2015d) for forecasts of gasoline prices and household gasoline expenditures. The EIA's estimate that households will save $700 on gas in 2015 is based in part on the EIA's projections of gas prices throughout the remainder of 2015. Although this report does not aim to re-estimate this number, we observe that on average individuals saved $22 per month on gas when comparing gas spending from Dec 2014–Feb 2015, when gas prices were at their trough, to one year prior, when gas prices were high. When projecting total spending on gas beyond just gas purchases made with Chase credit and debit cards, we find that the average person saved roughly $31 per month. Assuming gas prices remain constant for the remainder of 2015, this implies an average savings of $372 on an annual basis. Although this is somewhat lower than the $700 savings per household estimated by the EIA, our unit of analysis is the primary account holder, which reflects a mix of individuals and households.

2 This estimated saving of $600 represents the difference in interest expense in the first year of a 30-year fixed-rate mortgage of $120,000 associated with a 50-basis point decline in interest rates. Event studies estimate that the impact of the Federal Reserve Board's Large Scale Asset Purchases caused the 30-year fixed-rate mortgage interest rates to decline by roughly 20–100 basis points in the first round of asset purchases (QE1) starting in late 2008 (see, e.g., Gagnon et al. 2010 and Patrabansh et al. 2014).

3 For a discussion of this debate among governors, See Vock (2015).

4 The 2009 National Household Survey of Transportation shows similar demographic differences in terms of vehicle miles driven, and, in addition, that men drive more than women. See Santos et al. (2011) for a summary of transportation trends by demographic groups, and the Council of Economic Advisors (2015) for a discussion of how demographic trends are contributing to the decline in vehicle miles traveled.

5 There is substantial evidence from past oil price fluctuations that individuals spend more when gas prices decline and less as gas prices rise, and that the implied marginal propensity to consume is greater than one (Edelstein and Kilian, 2009; Hamilton, 2009; Edelstein and Kilian 2007). Edelstein and Kilian (2007) estimate that a 1% increase in energy prices translate into only a -0.04% change in discretionary income (given the share of income spent on energy) but a -0.15% change in real total consumption, implying a marginal propensity to consume well above one. As others have enumerated and shown, there are multiple pathways by which energy prices affect consumption: When gas prices fall, people have more discretionary income; they feel more optimistic about the economy and their personal finances, giving them the confidence to save less and purchase more durables; finally, they recognize that vehicles have lower operating costs and are thus more willing to purchase them. The most recent literature cited above provides evidence for the reverse effects when gas prices increase. There is contradicting evidence as to whether gas price increases and decreases impact the economy symmetrically. For example, Hamilton (2003) shows that oil price increases have a bigger impact on the economy than oil price decreases. Edelstein and Kilian (2007 and 2009) estimate impacts that are more comparable in size.

6 Gas price data for states are provided by GasBuddy.com.

7 The average incomes for individuals in each quintile of gas spending displayed in Figure 2 are $60,600 for quintile 1, $58,400 for quintile 2, $56,200 for quintile 3, $59,800 for quintile 4, and $70,000 for quintile 5.

8 In 2013, gas spending represented 4.1% of total income.

9 Each of the geographic analyses presented in this report (Figures 4 through 11) displays summary data aggregating credit and debit card transactions. The 23 states that have Chase branches have significantly higher proportions of debit card activity. For a more in-depth discussion of how this might influence our estimates, see the Data Asset and Methodology Section.

10 These regional differences contrast slightly with the 2014 Consumer Expenditure Survey that estimates gas spending to be the highest in the South and West ($213 per month per individual in both regions). We believe we may be underestimating gas spending in States in which we do not have a branch footprint. For example, in the West this includes Wyoming, Montana, and New Mexico. As a fraction of income, individuals spend the most on gas in the Midwest (4.3%) and South (4.7%), which is in line with our estimates. In addition, the 2014 CES finds gas spending to represent a larger fraction of income in rural areas (5.0%) compared to urban areas (3.4%). Similarly, the 2009 National Household Transportation Survey finds that households in less densely populated areas own more vehicles and drive more vehicle miles per year (Santos et al. 2015). Compared to employed individuals who live in large metropolitan areas, commuters who live in surrounding suburbs drive over 50% more miles, and commuters who live in rural areas and small towns drive almost twice as many miles (Perks and Raborn, 2013).

11 We assume here that everyone purchases gas in the state in which they live. State price data are provided by GasBuddy.com.

12 As recently reported by the EIA, supply disruptions can also cause temporary price shocks in certain markets leading to additional variation apart from gas taxes (EIA 2015c).

13 The implied price elasticity of demand of less than -0.30 is in line with recent research estimating the price elasticity of demand for gas at -0.37 (Coglianese et al. 2015).

14 The implied price elasticity of demand is more consistent with existing estimates of the price elasticity of demand used for modeling purposes by the EIA, which typically range between -0.02 and -0.04 (EIA 2015a).

15 This fraction (71%) represents our observed gas spending in 2014 as a fraction of the total per capita gas spending reported per consumer unit in the 2014 Consumer Expenditure Survey.

16 Our gas demographic findings are broadly consistent with national statistics. The 2014 Consumer Expenditure Survey (CES) finds gas spending to be highest among individuals aged 35–54 years in absolute terms, but individuals under 30 years old spend the highest fraction of their income on gas. In terms of income, the 2014 CES estimates monthly gas spending to be $97 or 11.3% of income for lowest quintile earners (less than $18,362 in income); $154 or 6.8% of income for second quintile earners ($18,362–$35,681); $203 or 5.2% of income for third quintile earners ($35,681–$59,549); $259 or 4.0% of income for fourth quintile earners ($59,549–$99,620); and $316 or 2.2% of income on gas top quintile earners (more than $99,620).

The income distribution of our population differs from the national population particularly at the low end of the spectrum, and gas spending estimates as reported by the CES are higher than JPMC Institute estimates particularly for high-income earners. In addition, data from the 2009 National Household Transportation Survey shows that women drive roughly one-third fewer vehicle miles than men on an annual basis, and that vehicle miles traveled peaks among individuals in their 30s and 40s and falls precipitously after age 60. We explore whether estimates reported may be biased by demographic characteristics of individuals who tend to be primary versus secondary account holders by calculating the same statistics for account holders who have a single authorized user on their account. We find similar results among this subsample of accounts with single authorized users.

17 For this analysis we assign people as low-gas spenders or high-gas spenders based on the average gas spending in each individual's zip code. See the Data Asset and Methodology section for a more in depth description of our approach.

18 Growth in non-gas card spend could be driven by not just economic growth, but also growth in Chase card usage relative to other payment mechanisms (e.g., cash, check and non-Chase credit cards).

19 As described in the Data Asset and Methodology section, we also explore whether the marginal propensity to consume differs by income group or by region. We find unreliable and therefore inconclusive results.

20 Non-gas spending increased by 5.9% for low-gas spenders, 5.2% for median-gas spenders, and 4.8% for high-gas spenders. These growth rates significantly higher than comparable estimates from the Census Bureau's Retail Trade Survey, from which we estimate 3.3% growth in per capita retail and food services spend not seasonally adjusted (and excluding auto and gas related spend).

21 These percentages are estimated based on observed levels of spending in 2014 in our data compared to those reported in the 2014 Consumer Expenditure Survey. See the Data Asset and Methodology section for a more detailed description of these estimates.

22 See for example Edelstein and Kilian (2007).

23 We also examine whether the marginal propensity to consume differed according to income group and region, but we do not find reliable results or discernible patterns. The results are described in the Data Asset and Methodology section.

24 Our findings are consistent with Gicheva et al. (2010) and Edelstein and Kilian (2007) who find that when gas prices increase, individuals reduce their restaurant expenditures. Both studies also find, however, that people increase their grocery expenditures overall since they are eating at home more, but that they switch to less expensive grocery purchases. Edelstien and Kilian (2007) provide evidence that gas price changes have a large impact on vehicle purchases and smaller but significant impacts on other durable goods, non-durable goods and certain services.

25 For example, we likely underestimate the impacts on certain services and durables, notably vehicle purchases, where debit or credit cards are not the most typical payment mechanism. Hamilton (2009) and Edelstein and Kilian (2007 and 2009) provide evidence that gas price fluctuations have a substantial impact on vehicle purchases.

26 It is worth noting that we do not observe itemized purchase receipts and therefore cannot distinguish between gas and convenience store purchases within gas stations. On the other hand, gas purchases at large discount stores are typically separate purchases and categorized as gas stations.

27 Research by the Federal Reserve Bank of San Francisco estimates that roughly 60% of total spend on food, personal care and general merchandise are made on credit or debit cards, compared to less than 50% for all other categories (Bennet et al. 2014).

28 When we assign individuals to gas spend quintiles based on their own gas spending in the High Price period (Dec 2013–Feb 2014), we observe that gas spending among top quintile gas spenders appears be lower in the years prior and after simply due to mean reversion. Shifting to the leave out mean decreases the spread in gas spending between low-gas and high-gas spenders. Average spending levels of median-gas spenders increase from $101 to $143 in gas spend and $1319 to $1432 in non-gas spend when we go from means to leave out means. The spread between low-gas and high-gas spenders also narrows when we shift to leave out means: low-gas spenders increase from $2 to $64 in gas spending and $1010 to $1432 in non-gas spend, and high-gas spenders drop from $359 to $196 in gas spending and $2290 to $1703.

29 As an alternative specification to adjust our estimates for pre-trends in percentage terms, we also calculated the difference in equations (3) and (4) relative to a counterfactual level, which assumes that gas spending and non-gas spending, respectively, had continued to increase during the treatment period at the same rate as they had during the pre-period. Mathematically, this is done by replacing $GasSpend_i^{HighPrice2}$ and $NonGasSpend_i^{HighPrice2}$ in equations (3) and (4) respectively with $GasSpend_i^{HighPrice2} * \frac{GasSpend_i^{HighPrice2}}{GasSpend_i^{PriorHighPrice}}$ and $nGasSpend_i^{HighPrice2} * \frac{NonGasSpend_i^{HighPrice2}}{NonGasSpend_i^{PriorHighPrice2}}$. This estimate removes the pre-trends in percent terms but still allows for a direct estimate of the MPC in dollar terms. This specification yielded very similar results: an MPC point estimate of 73% and a confidence interval of 44% to 103%.

30 We estimate the fraction of gas spend observed on card (71%) by dividing average monthly gas spend observed for Chase customers in 2014 ($146) by the average monthly consumer expenditure on gasoline and motor oil ($206) reported by the 2014 Consumer Expenditure Survey. Similarly we estimate the fraction of non-gas spend observed on card (58%) by dividing monthly non-gas card spend for Chase customers ($1,524) by the average monthly consumer expenditure on total non-gas consumption ($2,636) for 2014. In defining non-gas consumption within the Consumer Expenditure Survey, we exclude auto purchases, auto finance, shelter and pension related expenditures, which we believe are extremely unlikely to be expenditures made using credit cards. Although we believe benchmarking our estimates to the CES provides us with the best calibration, our results would have been qualitatively similar had we used other industry benchmarks.

Weathering Volatility

Big Data on the Financial Ups and Downs of U.S. Individuals

May 2015

JPMorgan Chase & Co.
INSTITUTE

About the Institute

The global economy has never been more complex, more interconnected, or faster moving. Yet economists, businesses, nonprofit leaders and policymakers have lacked access to real-time data and the analytic tools to provide a comprehensive perspective. The results – made painfully clear by the global financial crisis and its aftermath – have been unrealized potential, inequitable growth and preventable market failures.

The JPMorgan Chase Institute is harnessing the scale and scope of one of the world's leading firms to explain the global economy as it truly exists. The mission of the JPMorgan Chase Institute is to help decision-makers – policymakers, businesses and nonprofit leaders – appreciate the scale, granularity, diversity and interconnectedness of the global economic system and use better facts, real-time data and thoughtful analysis to make smarter decisions to advance global prosperity. Drawing on JPMorgan Chase's unique proprietary data, expertise and market access, the Institute develops analyses and insights on the inner workings of the global economy, frames critical problems and convenes stakeholders and leading thinkers.

The JPMorgan Chase Institute is a global think tank dedicated to delivering data-rich analyses and expert insights for the public good.

Acknowledgments

We would like to acknowledge Jamie Dimon, CEO of JPMorgan Chase & Co., for his vision and leadership in establishing the Institute and enabling the ongoing research agenda. Many others from across the firm – notably Peter Scher, Len Laufer, Max Neukirchen, Joyce Chang, Matt Zames, Judy Miller, Alexis Bataillon, Gordon Smith, Sally Durdan and Kristin Lemkau – have provided the Institute with the resources and support to pioneer a new approach to contribute to global economic analyses and insight.

We would also like to acknowledge the contribution of our fantastic team of research analysts and fellows, specifically David Wasser, Pascal Noel, Peter Ganong and Vijay Narasiman; and experts within JPMorgan Chase, including Bruce Kasman, Michael Feroli, Joseph Lupton, Jesse Edgerton and Colleen Briggs. This effort would not have been possible without the critical support of the JPMorgan Intelligent Solutions team of data experts, including Stella Ng, Steve Farrell, Joe Bimmerle, Tony Wimmer, Jay Galloway, Bill Bowlsbey and Michael Solovay; and, JPMorgan Chase Institute team members Rachel Pacheco and Kathryn Kulp.

Finally, we would like to acknowledge with gratitude the thoughtful and invaluable input of our academic advisors, including Michael Barr, Ray Boshara, Sendhil Mullainathan and Jonathan Parker. For their generosity of time, insight and support, we are deeply grateful.

Weathering Volatility

Big Data on the Financial Ups and Downs of U.S. Individuals

Diana Farrell
Fiona Greig

Dear Reader,

When it comes to the global economy, businesses, policymakers and nonprofit leaders look to the best information available to frame critical issues and determine how to address them most effectively.

It is becoming clear, however, that the best information available isn't always good enough. Hard data that describe the economy as it truly exists are hard to come by, making it difficult to have a complete understanding of the economy, how individuals and businesses make decisions and the reach of economic interconnectedness.

Instead of measuring granular transaction-level data, inferences are made from macroeconomic trends. Instead of observing changes in economic behavior, self-reported answers to survey questions drive analyses. As a result, economic policy has relied on inadequate or inaccurate information. And individuals, households, businesses and other organizations have felt the consequences.

It's time to use hard data and smart insights to address the complex problems that affect us all. That's why we established the JPMorgan Chase Institute.

By combining the power of big data with an increased understanding of how social science affects financial behavior, we have an opportunity to understand the economy as it *truly* exists – using observable data to provide an unprecedented level of detail. With our access to proprietary data, combined with thoughtful analysis from policymaking, academic and business experts, we can help decision-makers understand global economic shifts as they are happening, or even before they occur. As part of our mission, we'll convene leading economic minds to discuss insights, debate their implications, and draw actionable conclusions.

As the world economy has become more interrelated, it has become even more essential for us to connect the dots. I have spent my entire career using hard data to develop insights that address complex challenges, and I can attest that what we are doing here is truly unique. I am honored to have the opportunity to lead this new organization and, with our inaugural report, *Weathering Volatility*, deliver data-driven insights that, until now, would not have been possible.

We're excited to begin.

Sincerely,

DIANA FARRELL,
President and CEO, JPMorgan Chase Institute

Contents

Executive Summary

In this inaugural report, researchers from the JPMorgan Chase Institute analyzed proprietary data from JPMorgan Chase & Co. to determine how income and consumption fluctuate on a monthly and a yearly basis. Drawing from detailed transaction information for nearly 30 million customers, we constructed a unique data asset of 2.5 million account holders. We examined income and consumption habits on a transaction-by-transaction basis between October 2012 and December 2014 to draw conclusions about fluctuations in earning and spending among U.S. individuals.

THE JPMORGAN CHASE INSTITUTE DATA ASSET SAMPLE

27M
Chase accounts

from which a sample of
2.5M
is taken

100,000
people out of the 2.5 million sample were randomly selected and the following four types of information were analyzed

Criteria used to select the 2.5 million accounts include:

- Checking account in October 2012 to December 2014.
- At least $500 of deposits every month.
- At least five outflow transactions every month.
- Chase credit card in October 2012 to December 2014.

MONTHLY BALANCES
For 27 months on all Chase consumer products: **checking account, savings account, credit card, mortgage and home equity loans** and **auto loans**

135M TRANSACTIONS
Information on **amount, day and time, zip code, merchant** and **channel**

CREDIT BUREAU DATA
Estimate of **monthly payments** as well as current **outstanding balances** and delinquency statistics for credit cards, mortgages and other lines of credit

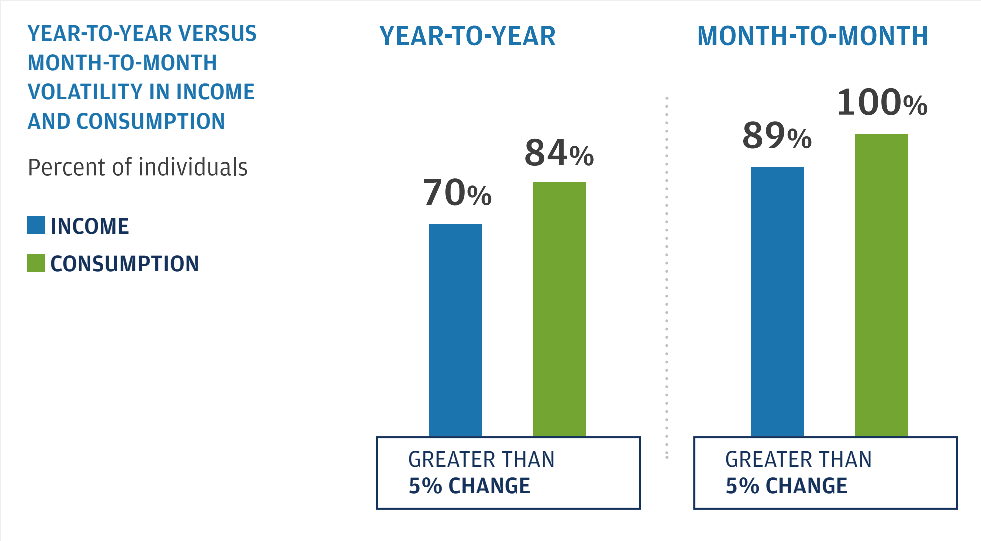
INDIVIDUAL CHARACTERISTICS
On an entirely de-identified sample: **gender, age** and **zip code**

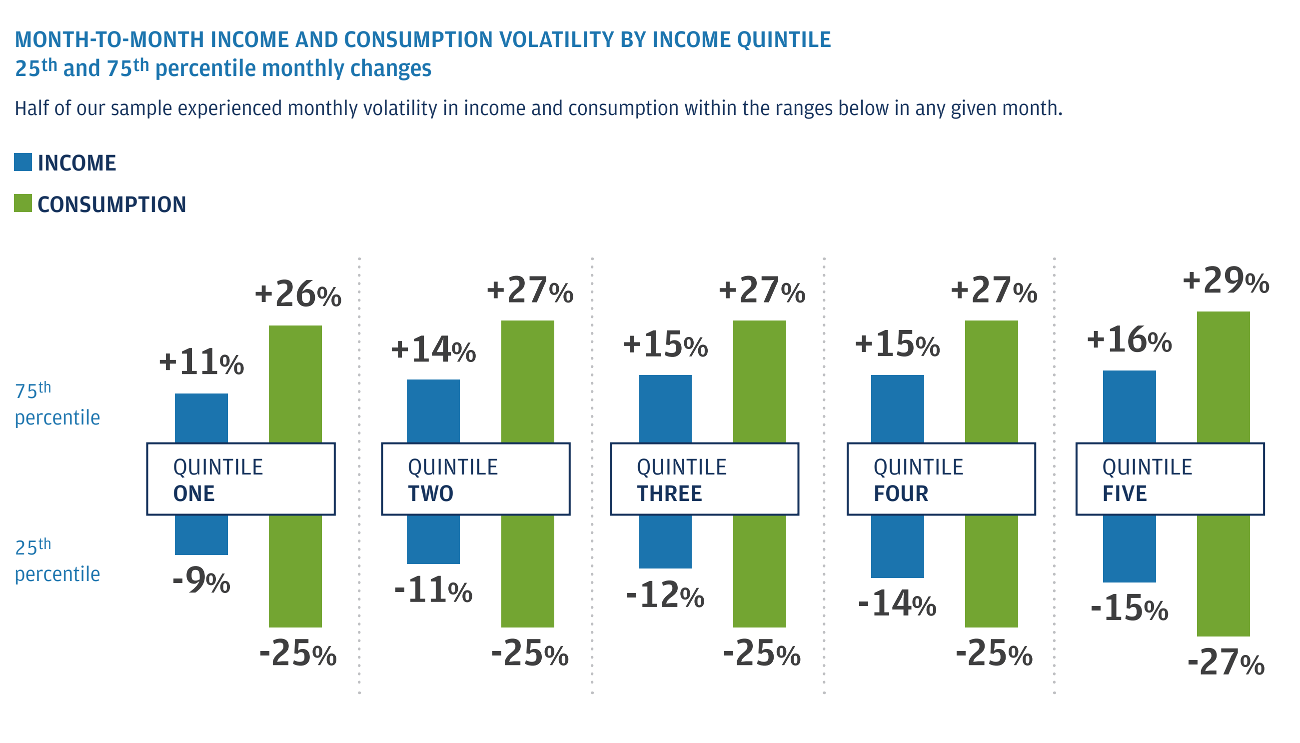
Our findings are summarized into three key points:

Finding One

Individuals experienced high levels of income volatility and higher levels of consumption volatility across the income spectrum.

Volatility was even greater on a month-to-month basis than on a year-to-year basis. Some of the drivers of monthly volatility included months with five Fridays, when individuals may be paid three times instead of two; tax bills and refunds; and the year-end shopping season.

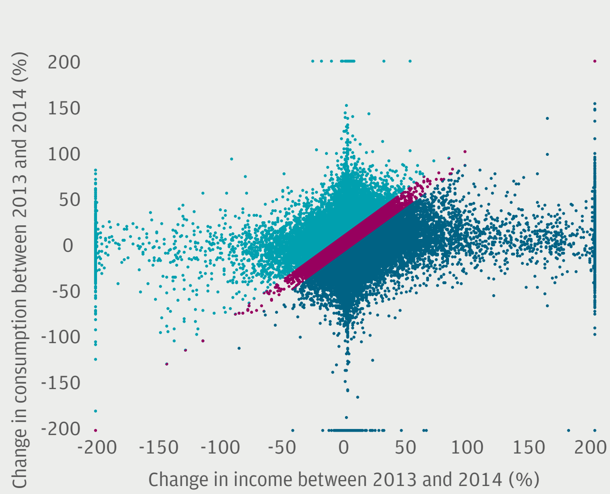




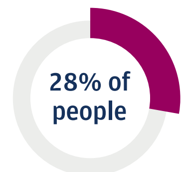
Finding Two

Income and consumption changes did not move in tandem; there was only a slightly positive correlation between changes in income and changes in consumption between 2013 and 2014. Three behavioral groupings describe the link between income and consumption changes:

BEHAVIORAL GROUPINGS OF INDIVIDUAL INCOME AND CONSUMPTION CHANGES



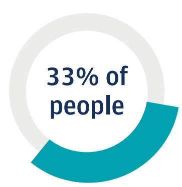
Responders



Individuals for whom income and consumption changes are within 10 percentage points of each other. Responders are more likely to have lower annual incomes and less access to liquidity through credit cards. They account for 28% of our sample.

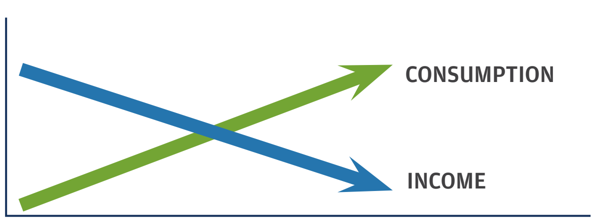
RESPONDERS:
Income and consumption changes are within 10 percentage points of each other.

Sticky **Optimists**

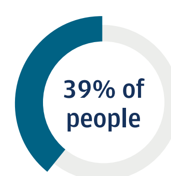


Individuals for whom consumption changes are higher than income changes by more than 10 percentage points. Sticky Optimists are more likely to have higher annual incomes and more spending power through credit cards. They account for 33% of our sample.

STICKY OPTIMISTS:
Consumption increases more than income by more than 10 percentage points.

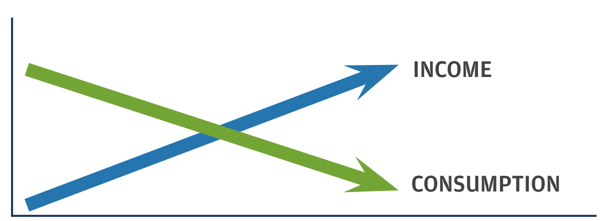


Sticky **Pessimists**



Individuals for whom consumption changes are lower than income changes by more than 10 percentage points. Sticky Pessimists are equally represented across income levels - from low-income to high-income - and they make up 39% of our sample.

STICKY PESSIMISTS:
Consumption increases less than income by more than 10 percentage points.

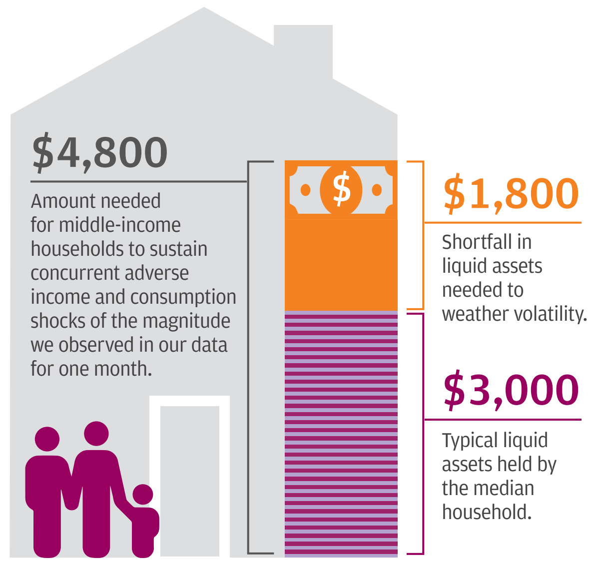


Finding Three

The typical household did not have a sufficient financial buffer to weather the degree of income and consumption volatility observed in our data.

The typical household did not maintain enough liquid savings that could be accessed immediately in the event of a large, unexpected expense sustained at the same time as a loss in income. While many in the field of consumer finance have long advised that consumers maintain an emergency fund, our research into income and consumption volatility shows that a financial buffer is a more important consideration for individuals across the entire income spectrum than is generally understood. We find that not only was volatility high for income and consumption, but also changes in income and consumption did not move in tandem. This creates the risk that people might experience a negative swing in income at the same time that they incur a large, potentially unexpected, expense. Based on our findings, we estimate that a typical middle-income household needed approximately $4,800 in liquid assets – roughly 14% of annual income after taxes – to have sustained the observed monthly fluctuations in income and spending but they had only $3,000. Required levels of liquid assets, however, were largely unavailable to most individuals across quintiles, except top earners.

LIQUID ASSETS NEEDED BY THE TYPICAL AMERICAN HOUSEHOLD TO WEATHER VOLATILITY



Conclusion

We conclude from these early findings that, given how noisy and unpredictable financial lives are, most individuals would benefit from innovative tools to better understand and manage their bottom line. These tools could include analytical platforms that help people track their earning and spending patterns as well as the sources, magnitude and timing of fluctuations in income and consumption. In addition, financial service providers, employers and policymakers can help individuals reduce and manage volatility, better match income and consumption changes or put these fluctuations to good use to help them save money. Potential solutions include new savings, insurance and credit products to help smooth income and spending; technical solutions, such as making deposited funds more immediately available to banking customers; and products or automated transfers that allow people to save during naturally occurring upswings in income, such as in five-Friday months and tax refund season.

Findings: Individual Income and Consumption Volatility

Most individuals in the U.S. are not prepared to sustain typical changes in their income or consumption. U.S. households do not have the necessary financial cushion to cover large expenses that may occur at the same time as a job loss or other reductions of income. These conclusions are based on a robust data asset assembled by the JPMorgan Chase Institute that shows changes in income and consumption, a lack of correlation between the two and the lack of liquid assets maintained by American individuals to weather a financial storm.

Finding One

Individuals experienced high levels of income volatility and higher levels of consumption volatility across the income spectrum.

Analysis of income and spending behaviors requires a robust set of data. The JPMorgan Chase Institute created a data asset encompassing a universe of 2.5 million customers. Though the individuals who make up the data asset differ from the nation in some important ways, they comprise a broad spectrum of individuals across income, age and geography. Using a random sample of 100,000 primary account holders (for the purpose of this report, "individual(s)" refers to those account holders comprised in the data asset), we categorized transactions into income, consumption and other activity to observe financial behavior.

We categorized individuals in our sample into five income quintiles and five consumption quintiles, ranging from the lowest income to the highest income and from the lowest consumption to the highest consumption. Doing so allowed us to examine how volatile income and consumption were within a given income quintile and also assess the degree to which individuals moved from one quintile to another.

While most existing academic and government research has focused on per capita income statistics of aggregate population data or limited surveys of individuals, we looked at the actual financial activity of individuals from month to month and observed income and consumption changes both in the aggregate and at the individual level. By taking a granular view over time, we observed the timing, magnitude and sources of income and consumption changes – both extreme and subtle.

Upon examining the data, we learned that: (1) individuals from across the income spectrum experience high levels of both income and consumption volatility, more so on a month-to-month basis than on a year-to-year basis; (2) income and consumption did not move together – for only a minority of the population did income and consumption move together; and (3) individuals needed a significant financial cushion – roughly \$4,800 among middle-income earners – to weather the degree of volatility in income and spending observed in our data. Yet, few individuals maintained this type of buffer, suggesting that volatility in income and consumption is an important consideration for individuals across the income spectrum, from low-income earners to high-income earners. These key findings contribute to the understanding of the financial lives of individuals in the United States. We describe them further in the following analyses.

Year-to-Year Income and Consumption Volatility

We find that individuals experienced significant year-to-year income volatility. In fact, only 30% of individuals experienced a change in income of 5% or less between 2013 and 2014 (see Figure 1).[1] Eighteen percent saw their income increase between 1% and 4% in that time frame, and 12% experienced a 0% to 4% drop in income. At the other extreme, 26% of individuals experienced a change in income of more than 30%, up or down, with most seeing increases. Thus, 44% of individuals experienced a 5% to 30% change in income year over year.

Individuals' spending fluctuated even more than individual income. Between 2013 and 2014, only 16% of our sample experienced less than a 5% change in consumption in either direction. At the other extreme, 14% of individuals increased consumption by more than 30% and 10% decreased consumption by more than 30%.

Comparing volatility in income and consumption, we find that individuals were almost twice as likely to experience "narrow" changes of less than 5% in income than in consumption. But they were much more likely to experience "large" changes of between 5% and 30% in consumption than in income. Specifically, 59% of individuals experienced consumption changes of between 5% and 30%, whereas only 44% experienced income changes of between 5% and 30%.

When it comes to big changes, a similar proportion of our sample experienced greater than 30% changes in income (26%) as the amount that experienced a greater than 30% change in consumption (24%).

Our data show a slight upward trend in both income and consumption between 2013 and 2014. Average income in our sample increased by 6.5%, while average consumption increased by 3.8%.[2] Although directionally consistent with national trends of a recovering economy, we do not infer macroeconomic trends from the JPMorgan Chase Institute data asset, which only covers 2013 to 2014.[3] Observing changes over a longer time period will reveal how persistent are the volatility and trends in income and consumption we see now. Income volatility has already been well documented in the economics literature, and our estimates of income volatility are within the range of what has been observed in the literature in year-over-year estimates.[4] We observed higher levels of income volatility within the year than existing sources. According to the Census Bureau's Survey of Income and Program Participation, the four-month standard deviation of changes in income was 44% in 2011 compared to 55% for our sample in 2013 and 51% in 2014.

FIGURE 1: DISTRIBUTION OF CHANGES IN INDIVIDUAL INCOME AND CONSUMPTION BETWEEN 2013 AND 2014

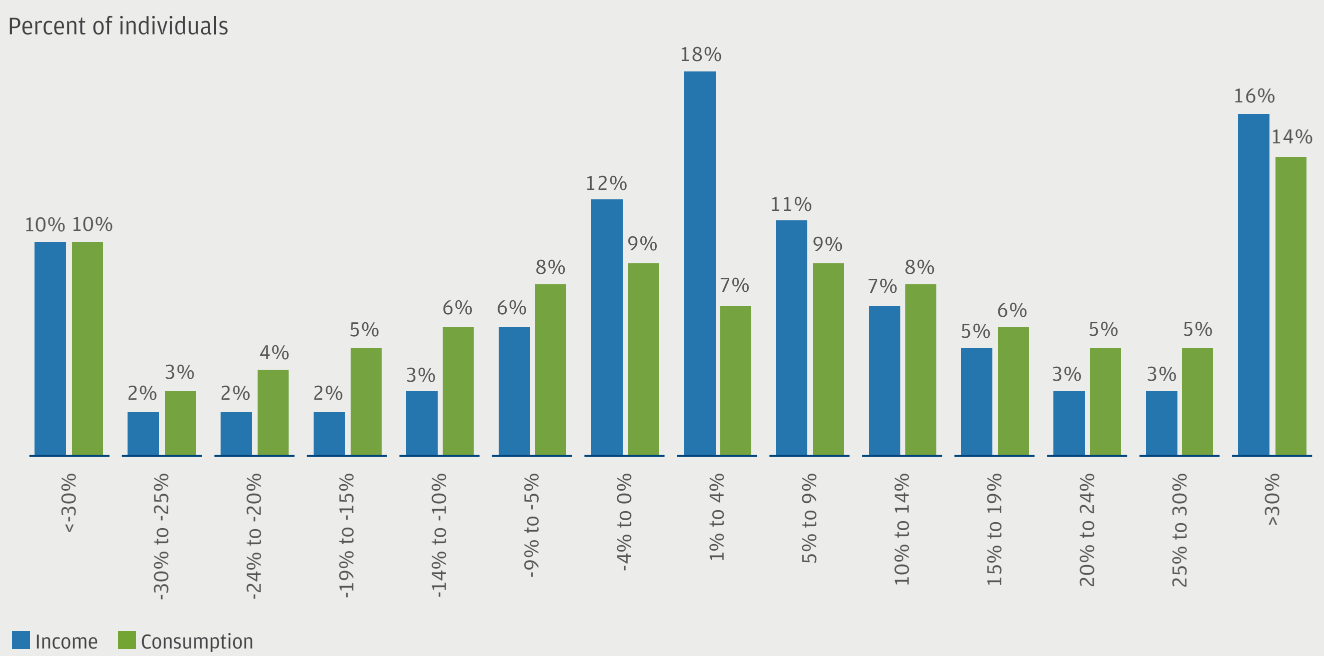


FIGURE 2: INCOME AND CONSUMPTION MOBILITY ACROSS QUINTILES (2013–2014)

Percentage of people transitioning across income quintiles between 2013 and 2014

2013 quintile \ 2014 quintile	1	2	3	4	5
5	2%	1%	2%	12%	82%
4	3%	4%	13%	67%	12%
3	5%	12%	65%	16%	2%
2	11%	68%	16%	3%	2%
1	78%	15%	3%	2%	2%

- 13% move down one or more income quintiles
- 72% remain in the same income quintile
- 14% move up one or more income quintiles

Percentage of people transitioning across consumption quintiles between 2013 and 2014

2013 quintile \ 2014 quintile	1	2	3	4	5
5	0%	1%	3%	17%	79%
4	1%	4%	21%	57%	17%
3	3%	21%	51%	21%	3%
2	19%	54%	22%	4%	1%
1	77%	19%	3%	1%	0%

- 18% move down one or more consumption quintiles
- 63% remain in the same consumption quintile
- 18% move up one or more consumption quintiles

Our results, however, make an important contribution to the understanding of consumption volatility. Prior literature on household consumption volatility is limited and typically measures the volatility of food consumption alone (both at home and in restaurants) based on the Panel Study of Income Dynamics (PSID).[5] Although we currently only compare two years of data, our data suggest that significant changes in consumption levels between years may be a wider phenomenon than previously understood. It remains unclear whether these fluctuations between years reflect true changes in lifestyles and welfare or simply the "lumpiness" of spending, as people pay for their new refrigerator, vacation, home repair or college tuition in one year but not the other. In any case it reflects the "sources and uses" reality of financial flows.

Mobility of Individuals Across Income and Consumption Quintiles

The volatility described earlier resulted in many individuals moving across income and consumption quintiles between 2013 and 2014. We observed greater consumption mobility than income mobility. Figure 2 shows the percentage of people who transitioned from one income quintile to another between 2013 and 2014.[6] Mobility between one income quintile and the next can represent a change in income of as much as $25,000. Based on Figure 2, for example, 78% of people who were in quintile 1 in 2013 remained in quintile 1 in 2014; 15% moved up to quintile 2 and 3% moved up to quintile 3.

Across our whole sample, 72% of individuals remained in the same income quintile between 2013 and 2014; the remaining 28% of the population moved up or down one or more quintiles.[7] This shows a level of income mobility that is consistent with mobility over much longer time periods as documented in other research (Debacker et al, 2012).

In terms of consumption mobility, the picture is notably less stable. Only 63% of the population remained in the same consumption quintile between 2013 and 2014. Consistent with the evidence presented above that consumption is more volatile than income, we also find more consumption mobility than income mobility between 2013 and 2014.

Month-to-Month Income and Consumption Volatility

Income and consumption volatility was higher on a monthly basis than on a yearly basis (see Figure 3 on page 9). While 70% of our sample experienced annual income changes of more than 5% between 2013 and 2014, on a monthly basis 89% of the sample experienced average monthly income changes more than 5% over the same time frame. Similarly 41% of individuals experienced fluctuations in income of more than 30% on a month-to-month basis compared to only 26% of people who experienced more than a 30% annual change in income between 2013 and 2014.

As with income, consumption volatility was greater when viewed at the monthly level: 84% of our sample experienced more than a 5% change in consumption over the course of a year, while 100% experienced more than a 5% change in monthly consumption over the same time frame.

Our data suggest that very few individuals follow a consistent monthly budget that sets strict parameters on spending. About 39% of individuals saw changes in consumption between 5% and 30%, and a full 60% of people experienced average monthly changes in consumption of greater than 30%.

Moreover, as shown in Figure 4 there was little correspondence in the timing of month-to-month changes in aggregate income and consumption. The month-to-month view suggests that many individuals experienced income and consumption movements simultaneously. As we discuss below, this raises the risk that unpredicted events can meaningfully affect an individual's financial stability.

Sources of Income and Consumption Volatility

There are a few notable sources of aggregate per capita income and consumption volatility, as depicted in Figures 5 and 6 (see page 10). Figure 5 demonstrates the considerable income volatility experienced over time and across all individuals by component. One can think of these sources as "seasonal" affect that might impact individuals more broadly.[8]

By far the largest source of identified income is labor income. Traditionally the steadfast backbone of an individual's liquidity, labor income was also the most volatile component of income. Some of the monthly labor income volatility can be attributed to December bonuses and to five-Friday months (November of 2012, March, May, August and November of 2013 and January, May, August and October of 2014). The average difference in labor income between a five-Friday month and the other months was 10%. Other drivers of labor income volatility included changes in hours worked, overtime wages and other factors not discernible in our data.

Other components of income were small in comparison to labor income and were generally more predictable. Tax refund season in February, March and April contributed to peaks in aggregate annual income in March and April, which is clearly evident in Figure 5 (see page 10). There was almost no volatility in Social Security and capital income (i.e., annuities and pensions) for the population in aggregate. Other income, which includes payments from other individuals and other miscellaneous or unclassifiable income, such as ATM cash deposits, was also stable and small.[9]

FIGURE 3: AVERAGE ABSOLUTE VALUE OF CHANGES IN INDIVIDUAL INCOME AND CONSUMPTION ON A YEARLY VERSUS A MONTHLY BASIS (2013–2014)

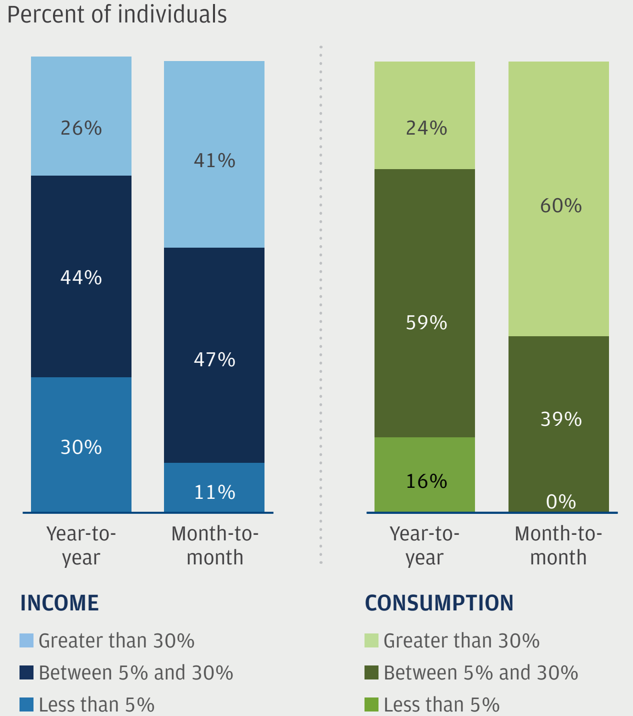


FIGURE 4: MONTH-TO-MONTH PERCENTAGE CHANGE IN INCOME AND CONSUMPTION

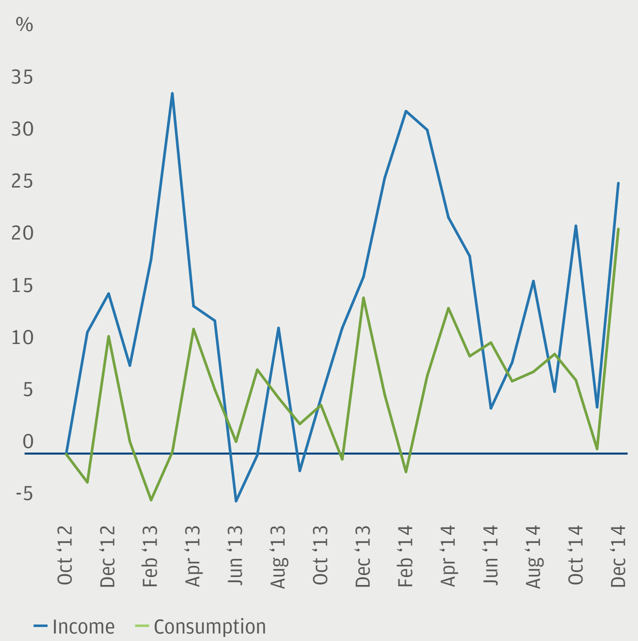


Figure 6 displays the key known components of consumption on a monthly per capita basis for all of the individuals in our sample. The largest and most volatile category of consumption was goods, such as groceries, household appliances and fuel, with spikes occurring around the end-of-year holiday shopping season. The next three largest categories of consumption were services; housing, including both rent and the non-principal portion of mortgage payments; and "other," including miscellaneous categories and outflows, such as ATM withdrawals. These three categories were less volatile than goods, though services appeared to mirror trends in the purchase of goods in a more attenuated way. Utility and debt payments were the next largest categories and remained fairly stable. Finally, payments to government, though the smallest spending category, spiked during tax season when many households made tax payments rather than received tax refunds.

Monthly Income and Consumption Volatility by Income Quintile

Top income quintile individuals experienced as much volatility in both income and consumption as bottom income quintile individuals. Figure 7 (see page 11) displays month-to-month volatility in income and consumption separately for each income quintile expressed in percentage terms.[10] The distribution of changes across the income quintiles suggests a widening spread from the lowest quintile to the highest quintile of income earners. We acknowledge that our estimates of volatility may be underestimated across the income spectrum but particularly in the lowest quintile because our sampling approach requires that individuals have a minimum of $500 in deposits each month.[11] Even so, comparing individuals across income quintiles 2 through 5 reveals comparable levels of monthly income volatility across the income spectrum. Individuals in the bottom income quintile experienced increases greater than 11% for 25% of the time and decreases greater than 9% for another 25% of the time. In comparison, a top quintile earner experienced income increases greater than 16% for 25% of the time and 15% drops in income for 25% of the time.

Consumption volatility is prevalent across the income ladder. The average person in the bottom two quintiles experienced a consumption increase of about 27% or decrease of 25% in half of the months.[12] Top quintile earners, similarly, saw consumption increase by 29% or decrease by 27% in half of the months.

The granular view of individual changes in income and consumption from one month to the next highlights how individuals across the income spectrum experienced dramatic volatility in income and consumption. What's particularly surprising is the degree of positive and negative fluctuation in both income and consumption. For individuals at all income levels, the degree of fluctuation was wider for consumption than for income.[13]

FIGURE 5: SOURCES OF VOLATILITY IN AGGREGATE PER CAPITA MONTHLY INCOME

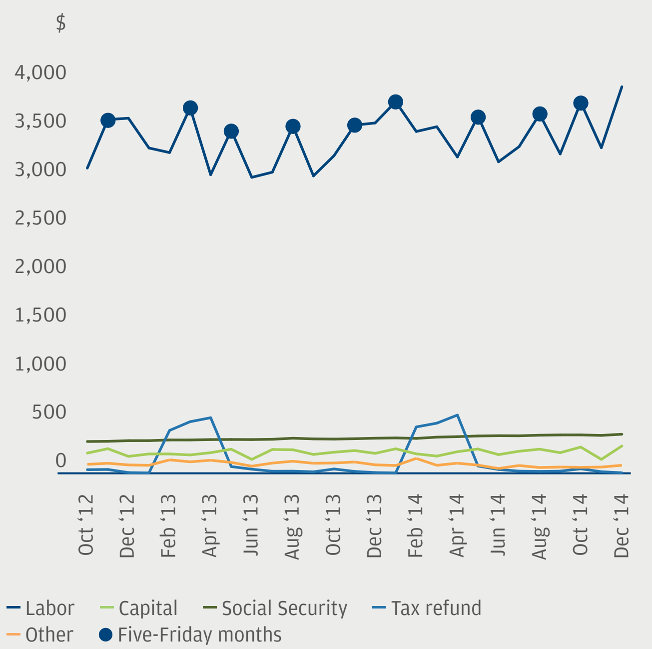
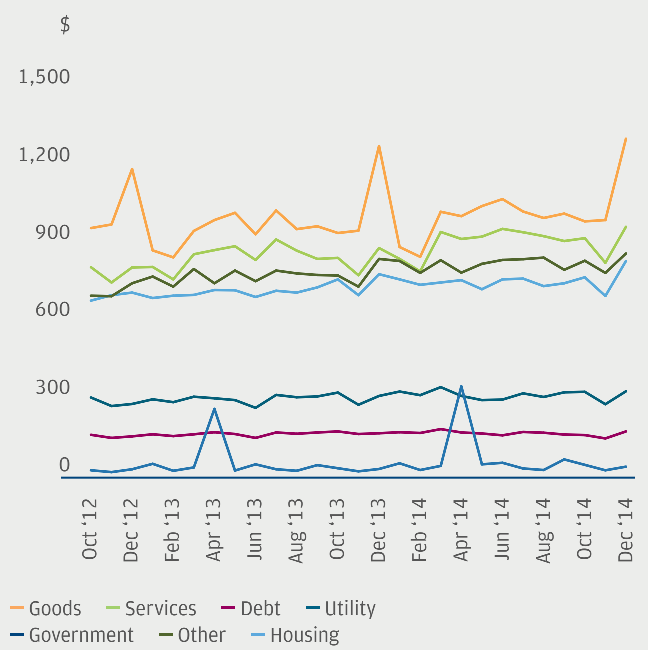


FIGURE 6: SOURCES OF VOLATILITY IN AGGREGATE PER CAPITA MONTHLY CONSUMPTION



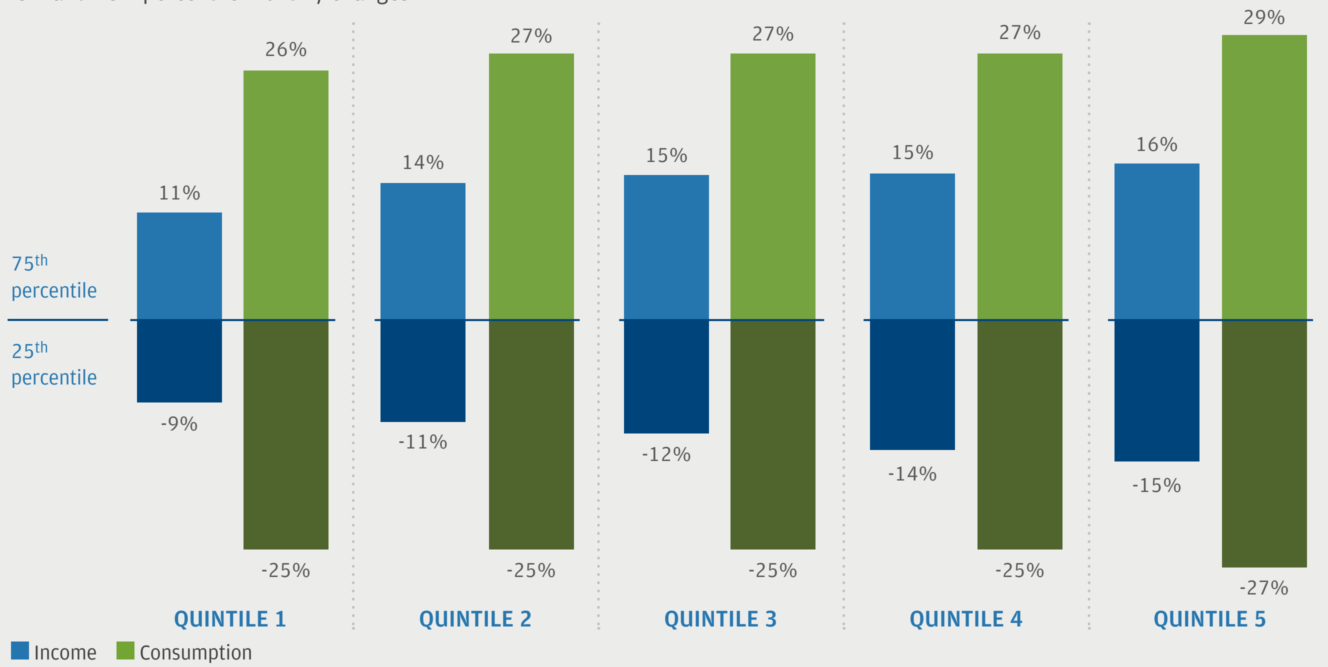
These are important findings. Scholars have long focused on the income volatility among low-income earners, both the extent of this volatility and its impact on people's ability to cover costs.[14] Our evidence suggests that no income group is immune to financial fluctuation – higher-income individuals experience as much volatility as lower-income individuals. Moreover, that individuals in every income quintile experience significantly more volatility in consumption than in income suggests that managing consumption shocks is critical to financial resilience. While some of these changes may be expected and predictable, other life events, such as sudden illness, are often unplanned and can disrupt stability, especially if the immediate cost far exceeds income. These wide swings in income and consumption can lead to instability at any level of income, highlighting the value of liquid assets to buffer against such shocks.

As discussed in Finding Three below, there is no one-size-fits-all liquidity balance that will serve as an appropriate buffer for individuals across incomes. The buffer required to weather income and consumption shocks is higher for high-income earners than for low-income earners given the higher levels of income.

We next turn to an equally important question for financial health: How, if at all, do income and consumption fluctuations move together? We find, generally speaking, that they do not.

"No income group is immune to financial fluctuation – higher-income individuals experience as much volatility as lower-income individuals."

FIGURE 7: MONTH-TO-MONTH INCOME AND CONSUMPTION VOLATILITY BY INCOME QUINTILE
25th and 75th percentile monthly changes



Finding Two

Income and consumption changes did not move in tandem; there was only a slightly positive correlation between changes in income and changes in consumption between 2013 and 2014. Three behavioral groupings describe the link between income and consumption changes.

Relationship Between Changes in Income and Consumption

Our data indicate only a very limited positive correlation between changes in income and changes in consumption. Figure 8 (below) plots our sample in terms of the percentage changes in income versus consumption that individuals experienced between 2013 and 2014[15]. The line represents the relationship between changes in income and changes in consumption.

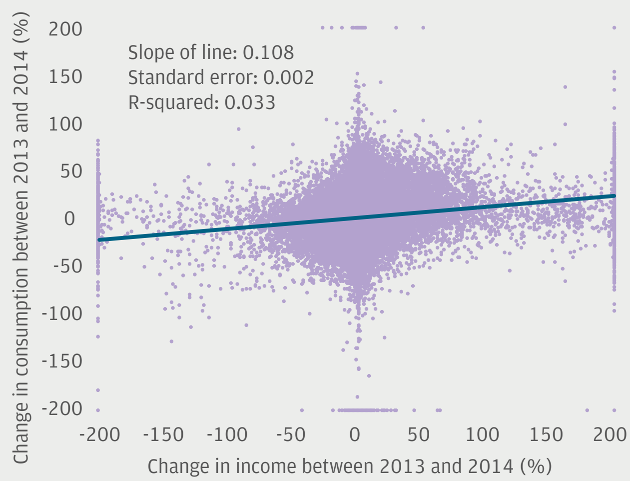
Figure 8 provides two new insights. First, points are scattered across the spectrum of income changes and consumption changes, with no strong discernible pattern between the two. Second, both income and consumption have trended upward, as evidenced by the many individuals in the top right quadrant of the chart who experienced a positive change in both.

It is important to note that the positive relationship between income and consumption changes is weak. The trend line overlaid on the chart indicates that for each 1% increase in income, individuals experienced a 0.1% increase in consumption. This relationship is statistically significant given our large sample size. However, given that these are changes over the course of two years, we might have expected a stronger relationship between changes in income and consumption.[16] In fact, 39% of people – everyone in the top-left and bottom-right quadrants of the graph – experienced changes in income and consumption that moved in opposite directions. In other words, they experienced either an increase in income while consumption decreased (24%); or, potentially more concerning, they experienced a decline in income while their consumption rose (15%).

We explore whether the relationship between income and consumption changes is sensitive to several economic and demographic factors. We find that the slope of the line is significantly steeper for individuals in income quintiles 1 and 2 (compared to those in quintiles 4 and 5), those who have outstanding balances on their credit cards (compared to those who do not), women (compared to men) and individuals aged 50 and older (compared to those younger than 50). Although statistically significant, these results are not economically significant in that the relationship between income and consumption changes remains only weakly positive even for these groups. Controlling for these variables explains less than one additional percent of the variance.

We also explored the relationship between an individual's income and consumption changes from month to month. We find an even weaker positive relationship between month-to-month changes in income and month-to-month changes in consumption, with a slope of 0.06 that does not explain even 1% of the variance.

FIGURE 8: SCATTER PLOT OF INDIVIDUAL CHANGES IN INCOME AND CONSUMPTION BETWEEN 2013 AND 2014



These findings have important implications. They suggest that individuals need to appreciate the degree to which income and consumption are volatile, and to prepare for the possibility that they might – unexpectedly or outside of their control – experience a negative swing in income concurrent with a positive swing in expenditures. Later in this section, we explore further the financial safeguards necessary to weather such swings concurrently.

Behavioral Groupings

Three behavioral groupings describe the link between income and consumption changes (see Figure 9 below).[15] These groupings are a first step toward understanding the prevalence of certain behaviors and how these different groups may react to future income and consumption shocks.

We describe the first group of individuals as "Responders." This group, which comprises 28% of our sample, consists of individuals for whom income and consumption changes fell within 10 percentage points of each other between 2013 and 2014. They respond to changes in income and consumption within a band of 10 percentage points.

A few characteristics distinguish this group. First, Responders are often responding to small and positive changes in income and consumption: 42% of Responders experienced less than a 5% change in income, and 34% saw less than a 5% change in consumption. Sixteen percent saw more than a 30% change in either income or consumption between 2013 and 2014. This suggests that small adjustments in income and consumption are generally easier to match and may not be the most important threat to an individual's financial stability.

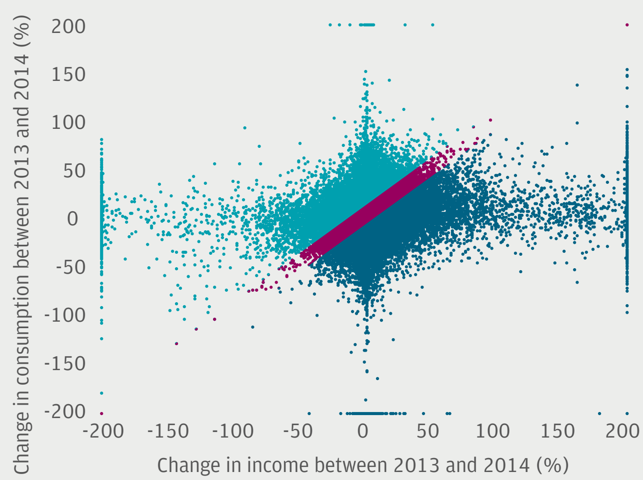
Next we sought to determine whether individuals who are more liquidity constrained – with lower income and/or less access to liquidity through credit cards – are less able to sustain a drop in income without a commensurate reduction in consumption. We use credit bureau data to estimate the degree to which individuals have already utilized credit cards as a source of liquidity.[17] Figure 10 suggests that Responders were slightly more likely to be among the bottom-income quintile and have more fully utilized their credit cards, giving them limited access to additional liquidity. Thirty percent of the bottom-income quintile earners were Responders compared to 25% of top quintile earners. Twenty-six percent of those who have no outstanding credit card balances were Responders compared to 32% of those who had outstanding credit card balances of more than 66% of their credit limit. In addition, Responders were more likely to be receiving Social Security. Thus, Responders appear to be a group who, possibly of necessity, either constrain consumption when they see a decrease in income or increase their earnings when they increase consumption.

FIGURE 9: BEHAVIORAL GROUPINGS OF INDIVIDUAL INCOME AND CONSUMPTION CHANGES

Responders 28%
Individuals for whom income and consumption changes are within 10 percentage points of each other between 2013 and 2014.

Sticky Optimists 33%
Individuals for whom consumption changes positively exceed income changes by more than 10 percentage points.

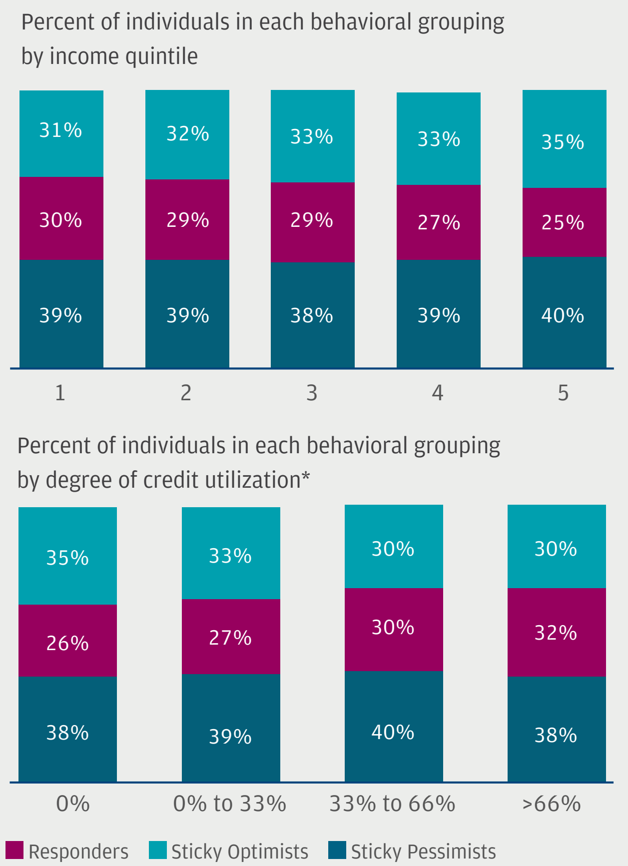
Sticky Pessimists 39%
Individuals for whom income changes positively exceed consumption changes by more than 10 percentage points.



Upon closer inspection of income and changes on a monthly basis, we found that virtually no one in this group matches their income and consumption changes month to month. In other words, even while typically responding to modest changes in income or consumption, Responders need more than one month to match changes in income with proportional changes in consumption, or vice versa. This implies that, in the month in which they experience an adverse fluctuation, they require liquid savings or access to credit to cover their expenses, which, as demonstrated above, they are less likely to have than our other two groups.

We describe the second behavioral grouping as "Sticky Optimists." Comprising 33% of our sample, these are the individuals whose consumption changes positively exceed their income changes by at least 10 percentage points. In other words, they maintain their spending level even when their income drops significantly; in a sense, they stick to their higher consumption pattern. Conversely, if they increase expenditures, either by choice or by necessity, their income does not increase commensurately. In the face of a positive change in income, they increase their spending by an even greater percentage. Sticky Optimists, on average, experienced a drop in income and a concurrent increase in consumption between 2013 and 2014: 67% experienced a drop in income of more than 5%. While 89% experienced an increase in consumption of more than 5%.

FIGURE 10: BEHAVIORAL GROUPINGS BY INCOME AND CREDIT UTILIZATION CHARACTERISTICS



* Credit utilization is measured as a person's total revolving balance on all credit cards as a percentage of their total credit limit across all credit cards.

Sticky Optimists are more likely to be higher earners. Figure 10 indicates that 35% of top quintile earners were Sticky Optimists, compared to 31% of bottom quintile earners. Figure 10, also indicates, as would be intuitive, that Sticky Optimists were more likely to have additional access to credit on their credit cards. Thirty-five percent of people who have no outstanding balance on their credit cards were Sticky Optimists, compared to 30% of people who had outstanding credit card balances of more than 66% of their credit limit. Although Sticky Optimists were more likely to have higher incomes and be less liquidity constrained, their response to income and consumption changes between 2013 and 2014 is unsustainable over the long term absent significant assets or additional income upon which they can draw.

We might expect individuals transitioning into retirement to be in this group, as their incomes potentially drop without a commensurate drop in consumption, but find that Sticky Optimists are less likely to be receiving Social Security payments and that the average age in this group is similar to the other two groups.

We describe the third behavioral grouping as "Sticky Pessimists." This group represents the largest number of individuals, at 39% of the sample. These are individuals for whom income changes positively exceed consumption changes by at least 10 percentage points. In other words, they stick to their lower consumption pattern despite an increase in income, and they drop consumption when income drops. In fact, many Sticky Pessimists experienced an increase in income and a drop in consumption at the same time. A full 66% saw income increases of more than 5%, and 65% saw consumption decreases of more than 5%. During the current income growth climate, these "cautious consumers" may be restraining unnecessary expenditures to even out any losses incurred during the economic downturn, or they may be maintaining consumption due to uncertain market expectations. Alternatively, these individuals may have more opportunity to increase income beyond their consumption needs or simply might have made a large one-time purchase in 2013 causing their 2014 spend to be lower than 2013. Interestingly, Sticky Pessimists were equally represented across the income and liquidity spectra, suggesting that these individuals are reluctant to spend outside their safety margin regardless of their level of income or access to credit card liquidity.

Finding Three

The typical individual did not have a sufficient financial buffer to weather the degree of income and consumption volatility that we observed in our data.

Low-income individuals are not alone in the degree of volatility in income and expenses they experience. Individuals across the income spectrum may face financial and liquidity management challenges. Our findings clearly underscore that individuals require a financial cushion to manage their cash flow as well as unexpected adverse swings in either income or spending. Based on evidence from the 2013 Survey of Consumer Finance on liquid assets, we find that the vast majority of U.S. households did not have sufficient liquid assets to cover the magnitude of volatility in both income and consumption evident in our data.

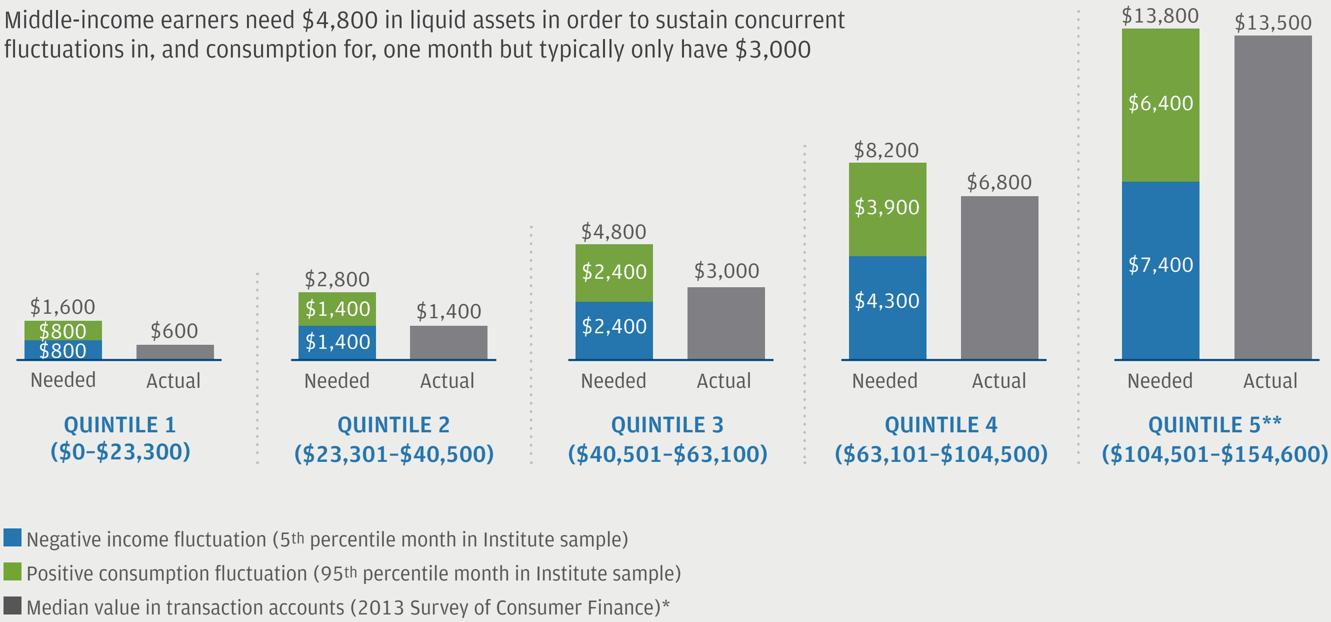
We use the month-to-month volatility observed at the individual level as illustrated in Figure 7 (see page 11) to estimate the amount of money individuals in each income quintile would have needed to safely absorb a negative fluctuation in income at the 5th percentile month, a positive fluctuation in consumption at the 95th percentile month, and both concurrently. We explore the level of assets needed to weather swings in both income and consumption, as we have observed that income and consumption swings do not move together either on a year-to-year or a month-to-month basis. Furthermore, volatility in consumption may reflect necessary lump sum payments (such as tuition payments) or changes in family circumstances, rather than in large discretionary outlays such as durable goods or vacations. Our estimates reflect the liquid assets that would have been required to sustain these adverse fluctuations for just one month. It may well be the case that adverse income and consumption changes, such as a job loss or a new medical condition, could persist for many months.

Although our population is somewhat more affluent than the general population, we apply the volatility estimates specific to each income quintile, as illustrated in Figure 7 (see page 11), to the income quintiles observed in the most recent Survey of Consumer Finance to provide rough estimates for the liquid assets needed to weather financial volatility across the population as a whole.[18] Specifically, we multiply the 5th percentile income change and 95th percentile consumption change by the median income for each income quintile as reported by the Survey of Consumer Finance. We then compare these levels of needed liquid assets to actual liquid assets as measured by the Survey of Consumer Finance for each income quintile.[19]

As shown in Figure 11 (see page 16), households in the bottom quintile needed a cushion of $800 in liquid assets to sustain 90% of the adverse income shocks observed in our data for one month. They needed an additional $800 to sustain a consumption shock for one month and a total of $1,600 to be able to sustain concurrent adverse income and consumption shocks of the magnitude we observe in our data. These liquid asset requirements increase significantly with each income quintile. For top quintile households this translates into a minimum of $7,400 to have sustained a negative income shock, $6,400 to have sustained a positive consumption shock, and $13,800 to have weathered income and consumption shocks of these sizes at the same time. The results in Figure 11 demonstrate how the degree of volatility evident in our data translates directly into the need for large financial cushions that increase in size for individuals with higher incomes.[20]

> *“Individuals need to prepare for the possibility that they might – unexpectedly or outside of their control – experience a negative swing in income concurrent with a positive swing in expenditures.”*

FIGURE 11: LIQUID ASSETS NEEDED FOR U.S. HOUSEHOLDS IN EACH INCOME QUINTILE TO WEATHER INCOME AND CONSUMPTION VOLATILITY FOR ONE MONTH (2013)



* Transaction accounts include checking, savings, and money market deposit accounts, money market funds, and call or cash accounts at brokerages including medical or health savings accounts and 529 education accounts.

** Quintile 5 reflects incomes for the 80th to 90th percentile.

Next to the liquid assets needed to withstand volatility, Figure 11 displays actual account balances for cash accounts maintained by each income quintile as reported by the 2013 Survey of Consumer Finance. Most households, except bottom income quintile households, had sufficient liquid assets to absorb 90% of negative income fluctuations. However, most households, except top income quintile households, did not have sufficient liquid assets to weather 90% of negative fluctuations in income and 90% of positive fluctuations in spending at the same time. An important conclusion from Figure 11 is that households in quintiles 2, 3 and 4 – the more typical U.S. households – had sufficient liquid assets to cover most fluctuations in income, but if they, for example, had needed to take a month of unpaid leave from their job and pay a large medical or tuition bill in the same month, they would have had difficulty doing so and would have likely needed to take on debt or liquidate other assets that are costly to access.[21] Low-income households would have had to do so even in the face a major negative swing in income for one month alone. Except for top earners, households across the income spectrum did not have sufficient liquid assets in place to weather 90% of financial fluctuations observed in our data. Even top quintile households might have struggled if faced with adverse shocks that persisted beyond one month. Thus, if faced with a big negative swing in income and a positive swing in expenses, most individuals would have likely been forced either to draw down on illiquid assets or to take on debt, both of which carry a price tag.

> *"Except for top earners, households across the income spectrum did not have sufficient liquid assets in place to weather the degree of financial fluctuations observed in our data."*

Implications for Individuals

Amidst so much volatility, understanding and managing one's financial bottom line is difficult.

The three findings above suggest a number of implications for individuals, financial institutions, employers and policymakers. First, given how noisy individual financial lives can be, individuals can benefit from tools to better understand and manage their bottom line. Understanding one's net income picture can be complex and difficult for the many individuals who cannot fully predict the changes they may face month to month and year to year. If many months out of the year are influenced by aberrations – from the mundane, such as five-Friday months, to the unexpected, such as the need to pay for a major home repair – it may be critical and difficult for individuals across the income spectrum to answer the most basic financial management question, such as:

- **Income:** What is my income in a typical month? By how much and when does my income fluctuate up and down?
- **Consumption:** What are my expenses in a typical month? What large one-time expenses do I have over the course of a year and when do I make them?
- **Cash management:** How much money should I have in reserve to weather fluctuations in my income and spending?
- **Bottom line:** Am I living beyond my means? How much money do I need in my rainy day fund to cover unexpected expenses and losses in income? Am I on track to meet my goal to save $5,000 for retirement this year?

Such basic questions are vexing precisely because income and expenditure can be so volatile. Getting a sense for one's bottom line requires a full accounting of not just consistent, recurring income and expenditures, such as regular paychecks and monthly expenditures on rent and groceries, but also anticipated but non-recurring income and expenditures, such as end-of-year bonuses and holiday spending, and unpredictable income and expenditures, such as a roof repair or job loss. Our measure of volatility in this report combines all identified sources of income and consumption.

The weak correlation between income and consumption changes suggests that people may be experiencing fluctuations in income and consumption that are unrelated. Moreover, the volatility in income and consumption could translate into balance sheet volatility: positively, in the case of the Sticky Pessimists, who saw larger changes in income than consumption, and negatively, in the case of Sticky Optimists, who saw smaller changes in income than consumption. This is less true for low-income earners, who were more likely to behave as Responders, suggesting that low-income earners increase earnings or cut discretionary or even non-discretionary spending when they experience shocks.

Importantly, the volatility that individuals experience carries not only a financial cost, but also psychological and cognitive costs. A recent study by Pew highlights that people favor financial stability over increasing income (The Pew Charitable Trusts, 2015b). Other researchers have previously shown that financial insecurity and scarcity exact a mental toll, making it more difficult for people to solve problems, exert self-discipline, and have the mental bandwidth to weigh the costs of borrowing or other short-term solutions.[22] While positioned as a problem that plagues low-income individuals, the scarcity caused by mismatched changes in income and consumption might be a more widespread experience than previously thought. We find that income and consumption volatility may be an important source of financial instability for individuals across the income spectrum, especially if their assets are small or illiquid. We demonstrate that the liquid assets of most U.S. households generally fall short of the levels required to cover the magnitude of most changes in income and consumption observed in our data.

"There may be an untapped opportunity for service providers, employers and policymakers to help individuals manage and mitigate financial volatility through innovative tools, products and programs."

Implications for Service Providers, Employers and Policymakers

Help Individuals Manage or Mitigate Volatility

From these early findings we conclude that there may be an untapped opportunity for service providers, employers and policymakers to help individuals manage and mitigate financial volatility through innovative tools, products and programs. These tools could include analytical financial planning platforms that integrate multiple aspects of a household's financial picture and help people see their typical earning and spending patterns and the sources, magnitude and timing of fluctuations in income and consumption. These tools may help people achieve not only better financial outcomes but also peace of mind.

There may be an untapped opportunity for financial products to assist individuals in "getting in front of" volatility and putting it to good use. The right financial tool could help individuals save (rather than spend) upswings in income, such as from five-Friday months, tax refunds or months with higher-than-typical earnings. For example, financial institutions could give individuals the option to automatically allocate to savings a specific dollar amount or percentage of income when their income exceeds a certain threshold or on predictable upswings such as five-Friday months or any tax refunds. Conversely, innovative insurance or credit products could also help individuals prepare for future unexpected dips in income or increases in necessary spending.

The magnitude and disparate timing of the income and spending fluctuations observed in our data suggest that people would benefit if they had real-time access to deposited funds in a way that was fully consistent with preventing fraud, currently only possible for same-institution deposits. According to the Federal Reserve Board's Diary of Consumer Payments, currently 46% of payment dollars are paid by check or electronic transfer, both of which require a minimum of one day before funds can be accessed by the payee (Bennett et al, 2014). Only a few transaction channels allow funds to be transferred and accessed by the payee immediately. These include wire transfers and, more recently, general-purpose immediate fund transfers, pioneered in the United States primarily by non-bank financial institutions.[23] Our research suggests that work currently under way by the Federal Reserve Board and financial institutions to improve the U.S. payment system by, for example, enabling same-day Automatic Clearing House electronic transfers could be a valuable step forward for individuals who do not have the financial buffer estimated in this report to be necessary to cover typical swings in income and spending.

Opportunities to help individuals mitigate or better match income and consumption volatility also extend to the workplace and public policy. Employers may want to consider more consistent work schedules as well as pay cycles and structures that better match consumption needs. These could include opting to pay employees on the first and 15th of every month (with amounts paid calibrated to reflect the length of the month) rather than every two weeks to better match payroll with large monthly outlays, such as rent, mortgage and other loan payments. Other workplace benefits, such as emergency funds, could help insure employees against the financial shocks they experience in their lives, which can reduce productivity and easily disrupt their ability to work.

Governments may want to pay out tax refunds more gradually or, if taxes are owed, structure and aggressively promote payment plans that allow individuals to smooth their payments in advance and after tax time. In the absence of more gradual payment mechanisms, financial institutions, policymakers and nonprofits could create more innovative products and services that assist individuals in saving their tax refunds or saving money in advance of tax payments.

In conclusion, managing volatility in income and in consumption looms large across all income quintiles to a greater extent than is generally understood. Total financial volatility for a given individual is potentially even higher, as changes across income and consumption do not move in tandem. Liquidity buffers that would help individuals weather typical volatility can represent a very large percentage of average incomes, constituting liquidity levels largely unavailable to most individuals. Better tools to help individuals understand and better manage their bottom line amidst these financial fluctuations are needed across the income spectrum, as are measures to increase predictability in income and consumption and match income to expenditure over time.

The JPMorgan Chase Institute Data Asset

In this report, the JPMorgan Chase Institute seeks to inform the public debate on the financial lives of U.S. individuals. To draw conclusions about household liquidity and income and consumption volatility, we adapted the firm's internal data on nearly 30 million U.S. account holders into a secure groundbreaking data asset. As the first financial institution to channel this wealth of information for the benefit of the public good, JPMorgan Chase put strong guardrails and strict privacy protocols in place to protect personal information throughout the creation and analysis of this data asset.

Data Privacy

The JPMorgan Chase Institute has adopted rigorous security protocols and checks and balances to ensure all customer data are kept confidential and secure. Our strict protocols are informed by statistical standards employed by government agencies and our work with technology, data privacy and security experts who are helping us maintain industry-leading standards.

There are several key steps the Institute takes to ensure customer data are safe, secure and anonymous:

- Before the Institute receives the data, all unique identifiable information – including names, account numbers, addresses, dates of birth and Social Security numbers – is removed.
- The Institute has put in place privacy protocols for its researchers, including requiring them to undergo rigorous background checks and enter into strict confidentiality agreements. Researchers are contractually obligated to use the data solely for approved research, and are contractually obligated not to re-identify any individual represented in the data.
- The Institute does not allow the publication of any information about an individual consumer or business. Any data point included in any publication based on the Institute's data may only reflect aggregate information.
- The data are stored on a secure server and can be accessed only under strict security procedures. The data cannot be exported outside of JPMorgan Chase's systems. The data are stored on systems that prevent them from being exported to other drives or sent to outside email addresses. These systems comply with all JPMorgan Chase Information Technology Risk Management requirements for the monitoring and security of data.

The Institute provides valuable insights to policymakers, businesses and nonprofit leaders. But these insights cannot come at the expense of consumer privacy. We take every precaution to ensure the confidence and security of our account holders' private information.

The Institute's data asset and research complement a giant body of surveys and other tools used to understand the financial behavior of individuals and businesses in the United States. Traditionally, research on earning, spending and financial behavior has relied primarily on a number of recurring public surveys in which individuals or establishments self-report their income, expenses or business sales. These surveys are costly to administer and often experience low response rates that recently have been falling even lower.[24] Public agencies typically administer these surveys periodically, seldom more than once a year, and sample 4,000 to 60,000 individuals (in the case of the Census). Typically these surveys gather information on only a few dimensions of financial behavior – either income or consumption, but not both. Private research organizations and think tanks also conduct a number of important and insightful surveys focused on financial health issues. Recent examples include the U.S. Financial Diaries conducted by the Center for Financial Services Innovation (CFSI) and NYU's Wagner's Financial Access Initiative (FAI) in 2013 and the Survey of American Family Finances conducted by The Pew Charitable Trusts in 2014.[25] Figure 12 provides an overview of the most common public, recurring surveys.

In 2014, spending on gas peaked on Friday, May 23, the Friday of Memorial Day weekend, and fell by 75% to a low point on December 31, New Year's Eve.

People spend three times as much on Mondays, the highest spending day of the week, as they do on Sundays, the lowest spending day of the week.

FIGURE 12: RECURRING SOURCES OF PUBLIC DATA ON THE FINANCIAL LIVES OF U.S. HOUSEHOLDS

Source	Data	Description	Sampling Approach and Size	Response Rate	Frequency
CENSUS	Current Population Survey	Personal income, labor force statistics	60,000 housing units from 824 sample areas	90%	Monthly
	Survey of Income and Program Participation	Personal income, income volatility, economic well-being, asset ownership, health insurance, housing expenditures	National panels: 14,000 to 52,000 households	70%	2.5 to 4 years
	Retail Trade and Food Services Survey	Personal consumption, sales/ inventories at/held by retail, ecommerce, food stores	12,000 to 22,000 retail businesses with paid employees	60% to 80%	Monthly/ annually
BUREAU OF ECONOMIC ANALYSIS	National Income and Product Accounts	GDP, personal income, savings, fixed Investment	National/aggregate of various government surveys	N/A	Monthly
BUREAU OF LABOR STATISTICS	Consumer Expenditures Survey	Personal/consumer unit consumption and income	Nationally representative sample of 7,000 consumer units for two one-week diaries and four interviews quarterly	72% to 75%	Monthly
FEDERAL RESERVE	Survey of Consumer Finance	Family income, net worth and asset and debt holdings	Nationally representative sample of 6,026 families	60%	Every three years
	Survey of Household Economics and Decisionmaking	Personal finances of households, credit access and behavior, student debt, savings, retirement and health-related expenses	Nationally representative sample of 4,134 households	70%	N/A (New)
UNIVERSITY OF MICHIGAN	Panel Study of Income Dynamics	Longitudinal study of 5,511 families on economic, education, health and financial outcomes	Nationally representative sample of family members of an original sample of 5,511 households	94%	Every two years
IRS	Tax return data	Annual income tax return data on tax filers	National	100%	Annually

More recently, scholars are turning to selected government records, including comprehensive IRS tax filings of individuals and their dependents, and records from participants in government programs such as Medicare or Social Security.[26] These administrative data have the benefit of offering large samples that are more likely to represent individuals' actual financial behavior than what is reported on surveys, but they lack a comprehensive and integrated view of both income and consumption. On a more limited basis, large-scale private data sets also are becoming available to researchers. These data sets include public data aggregators, such as Zillow.com and Redfin; personal financial websites that aggregate daily transactions and/or financial accounts; and other online tools. These sources offer a window into real behavior on a high-frequency basis. In particular, information from personal finance websites also has the potential to provide a near-complete picture of an individual's financial life if account holders are sufficiently active users.

The Institute's data asset combines access to daily account data with the ability to track the same account holders over time, creating a unique data asset that is comprehensive and consistent. The data asset differs from existing data sets in a number of important ways that help to make new contributions to the general understanding of the way individuals manage their money. First, our large sample of roughly 2.5 million individuals enables us to make observations of a broad and diverse population as well as focus on interesting subpopulations, such as retirees or other demographic groups. Second, our data are based on actual behavior of the same individuals over time with low attrition from month to month, offering a longitudinal, dynamic perspective rather than one-time snapshots. The data asset also offers a window into both inflows and outflows of financial accounts, complemented with credit bureau data on liabilities, offering a more complete perspective of earning and spending. Finally, unlike personal finance websites, which typically rely exclusively on transaction text descriptions to categorize transactions, our data include significant information on each transaction, including: merchant information for all debit and credit card purchases; the transaction channel by which the funds flowed; and a significantly longer text string for all electronic transfers that includes important payee and payer identification numbers that enhance our categorization algorithm. In short, this new data asset offers granular, high-frequency, longitudinal data on multiple dimensions of financial behavior.

Constructing our Sample

In constructing our data asset, we sought to provide an integrated profile of the financial lives of individuals. For the purposes of this research, the unit of analysis is the primary account holder, whom we subsequently refer to as individuals.[27] To avoid double counting financial activity, all joint accounts are captured under one individual, the primary account holder.

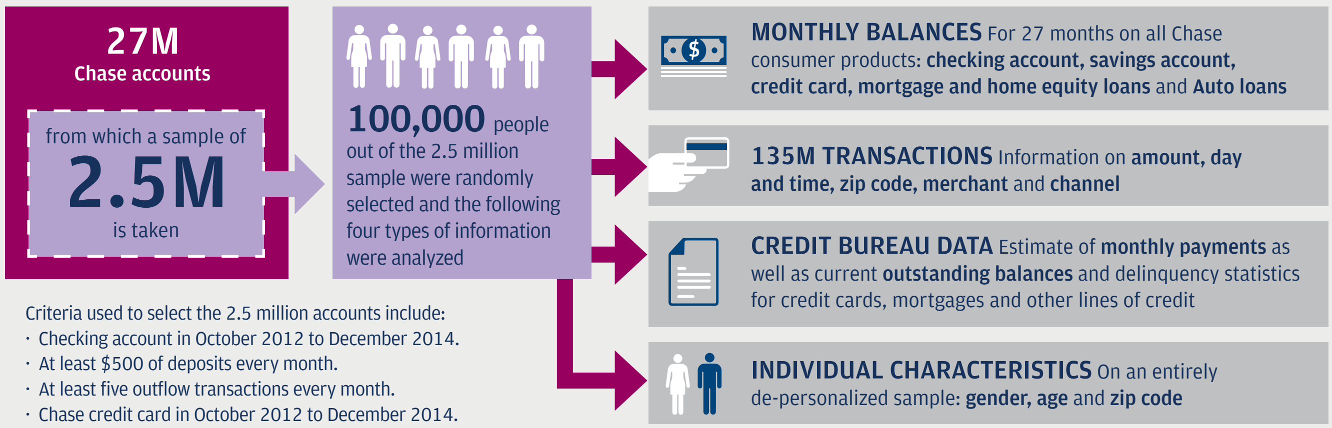
From almost 30 million accounts, we created a subsample of 2.5 million individuals for whom we have a near-complete view of their finances. To do so, from our initial universe of account holders, we selected individuals who maintained an active checking account every month between October 2012 and December 2014 with a monthly minimum of $500 in deposits and at least five outflows. These active users are considered to be "core" customers of the bank. In addition, we selected only individuals who kept an open Chase credit card for all 27 months, allowing us to analyze additional financial information reported by other banks to the credit bureaus. Applying these criteria, we culled our subsample of 2.5 million individual account holders, from which we drew a random sample of 100,000 individuals for use in this report.

The single highest spend day in 2014 (excluding Tax Day) was Monday, March 3, the Monday after the single highest earning day, which was Friday, February 28.

In aggregate our sample earns roughly 10% more in "Five-Friday" months.

The spring of 2014 was a celebratory time. The top 12 days for spending on restaurants and bars occurred between February 14, 2014 and May 11, 2014.

FIGURE 13: THE JPMORGAN CHASE INSTITUTE DATA ASSET SAMPLE



We have four types of depersonalized data for each individual:

1. **Monthly balances:** Monthly balances for all consumer products used by primary account holders in our sample, except for co-branded credit cards (for example, a merchant credit card issued by Chase). These products include checking accounts, savings accounts, certified deposit accounts, and monthly payments and outstanding balances on Chase borrowing products, such as credit card, mortgages, home equity loans and auto loans.
2. **Transactions:** Record of all inflow and outflow transactions that take place out of the checking account (including debit card and credit card transactions).
3. **Credit bureau data:** Estimate of monthly payments as well as current outstanding balances and delinquency statistics for credit cards, mortgages and other lines of credit.
4. **Individual characteristics:** Characteristics, such as age, gender and zip code. In addition, independent of the Institute's estimates of individual income derived by categorizing account inflows, for each individual, JPMorgan Chase calculates an estimated pre-tax annual income based on individual, third-party and zip code information. As described in the Findings: Individual Income and Consumption Volatility section, we use these data in specific analyses; for example, when we construct and segment our sample by income quintile.

Our sample of 100,000 people is different from the nation in a few important ways. First, our sample is biased geographically by Chase's physical branch footprint, which only covers 23 states.[28] Figure 14 compares the share of individuals in our sample in each Census region to the share of the total U.S. population in each region (according to the Census) and to the share of the banked population in each region (according to the FDIC Survey of the Unbanked in 2013). Our sample gives us broad coverage of the four Census regions, but with a bias in favor of the Northeast.

FIGURE 14: REGIONAL DISTRIBUTION OF THE INSTITUTE SAMPLE COMPARED TO THE NATION

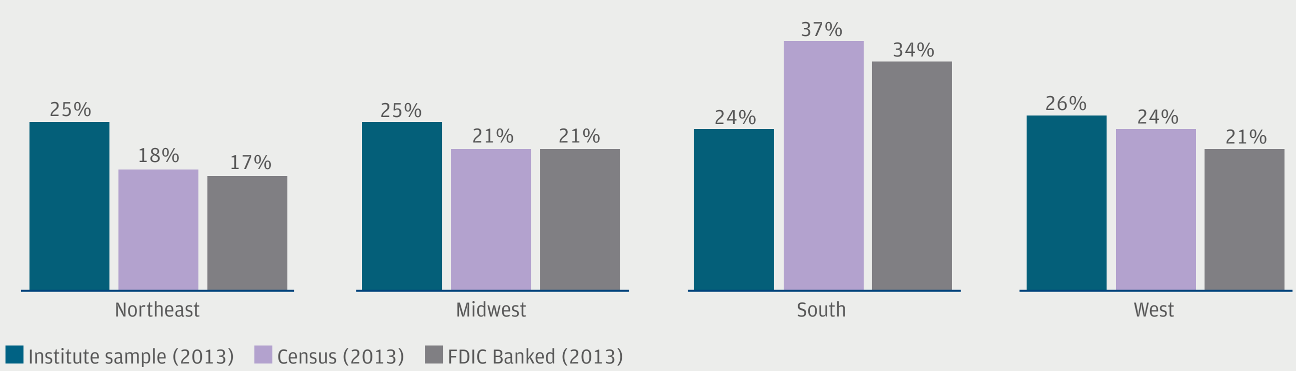
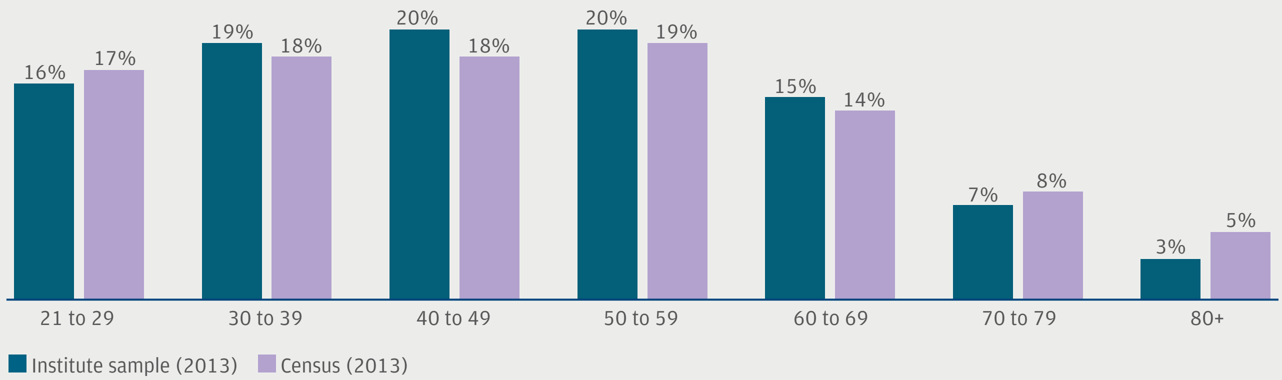
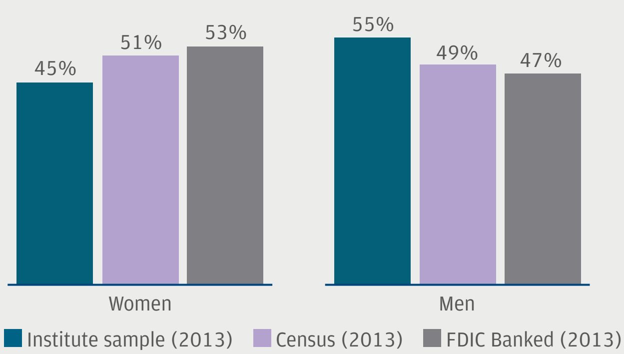


FIGURE 15: AGE DISTRIBUTION OF THE INSTITUTE SAMPLE RELATIVE TO THE NATION*



* Age distribution reflects a population over 21 years.

FIGURE 16: GENDER DISTRIBUTION OF THE INSTITUTE SAMPLE RELATIVE TO THE NATION



Second, as shown in Figure 16, our sample is skewed in favor of male account holders: 55% of our sample is male (compared to 49% for the 2013 Census and 47% for FDIC Banked), and 45% is female (compared to 51% for the 2013 Census and 53% for FDIC Banked).[29] This bias may reflect a tendency for men to be listed as primary account holders on joint accounts rather than an underlying bias in the Chase population in favor of men. Our sample is comparable to the nation in terms of average age but slightly underrepresents individuals aged 21 to 29 and aged 70 and above compared to the nation.

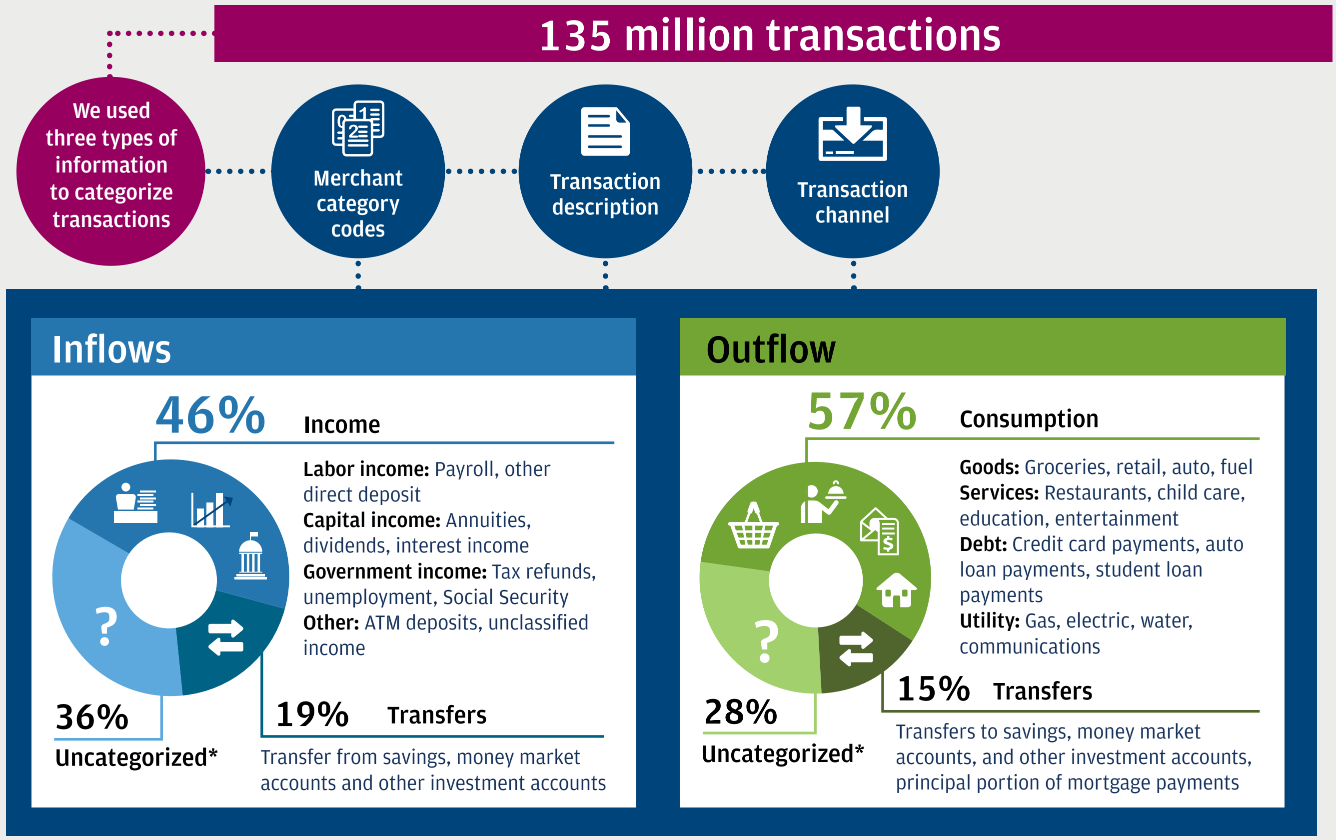
Finally, our sample is skewed in favor of higher-income individuals for a number of reasons. In our data asset, we observe only those individuals who have a relationship with Chase. Roughly 8% of Americans do not bank with a U.S. financial institution and tend to be disproportionally lower income and non-Asian minorities (FDIC 2014). In addition, our sampling criteria bias our sample in favor of higher-income individuals within the universe of Chase customers. The lowest income earners may not meet the sampling criteria of having $500 in deposits every month. And, because Chase does not operate in the subprime credit market, Chase credit card holders have credit scores above a specific threshold. Thus, Chase credit card holders skew towards higher-income earners.

Making Sense of the Data

On average, the individuals in our sample saw more than $8,000 moving in and out of their accounts each month, of which a significant portion represented transfers to and from other Chase and non-Chase financial accounts. Yet, money coming into an account cannot immediately be classified as income, nor can money moving out of an account be immediately classified as consumption. Through a number of techniques, we separate inflows into actual income and "dis-saving," or transfers from other financial accounts. Similarly, we separate outflows into consumption and saving. Figure 17 (on page 24) provides an overview of the outcome of our classification.

We use several strategies to categorize incoming and outgoing transactions into income, consumption and other categories. Specifically, we analyze merchant information to accurately sort debit and credit card purchases into appropriate consumption categories, such as grocery, fuel or department store. For electronic transfers, we categorize transactions into, for example, mortgage or utility payments. We also exploit the transaction channel by which the funds flow to categorize inflow and outflow transactions when payee or merchant information are not available. For example, we assume that all ATM cash withdrawals represent consumption and all ATM cash deposits represent income.

FIGURE 17: CATEGORIES ASSIGNED TO INFLOW AND OUTFLOW TRANSACTIONS



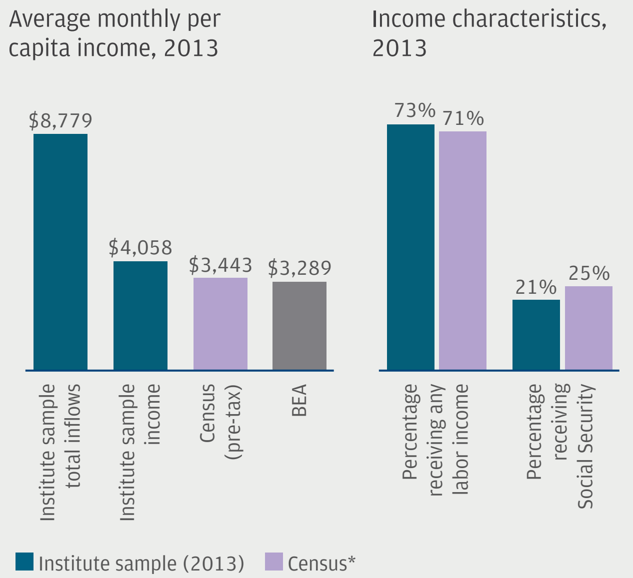
* Certain types of transactions, such as check deposits or check withdrawals, were unknown and thus remained uncategorized.

We are still left with some unidentified transactions – 36% of inflow dollars and 28% of outflow dollars – that are not included in this report's analyses. Despite these unidentified transactions, the resulting mean income levels in the Institute data asset are higher than national averages (see Figure 18 on page 25). Total average monthly inflows for the sample are $8,779, of which we identify $4,058 (46%) as income. This income figure is higher than estimates of $3,289 from the Bureau of Economic Analysis' National Income and Product Accounts which show monthly per capita disposable income (after taxes), as well as the Census Bureau's Current Population Survey at $3,443, which measures individual income before taxes.[30] In addition, we find that 73% of our sample received some form of labor income in 2013 compared to 71% nationally, and 21% of our sample received Social Security payments in 2013 compared to 25% nationally.

Figure 19 (on page 25) also indicates that we find higher levels of consumption than national estimates. We observe average monthly outflows of $8,247, of which $4,690 (57%) is consumption. In addition, we complement our Chase data with credit bureau data that allow us to identify additional consumption that may or may not be flowing through the Chase account.[31] The credit bureau data show an additional $1,200 worth of consumption, leading to much higher levels compared to the other sources. The consumption levels in the Institute sample, both with and without credit bureau data, are higher than national estimates from the Bureau of Economic Analysis's National Income (at $3,021) and Product Accounts and the Bureau of Labor Statistics' Consumer Expenditure Survey (at $4,258).[32]

The income and consumption statistics shown in Figures 18 and 19 confirm that the Institute sample is biased in favor of individuals who earn and spend more than the average individual nationally. Moreover, they reveal that we have been able to identify proportionally more outflows as consumption than inflows as income. As a result, our consumption estimates exceed our income estimates. These comparative statistics underscore the fact that the focus of this report is on the dynamic changes in income and consumption rather than the absolute levels of income and consumption. We examine the volatility of income and consumption and how they change in relation to one another. To further emphasize this point and more accurately highlight changes within the income spectrum, our findings are also shown by individual income quintile.

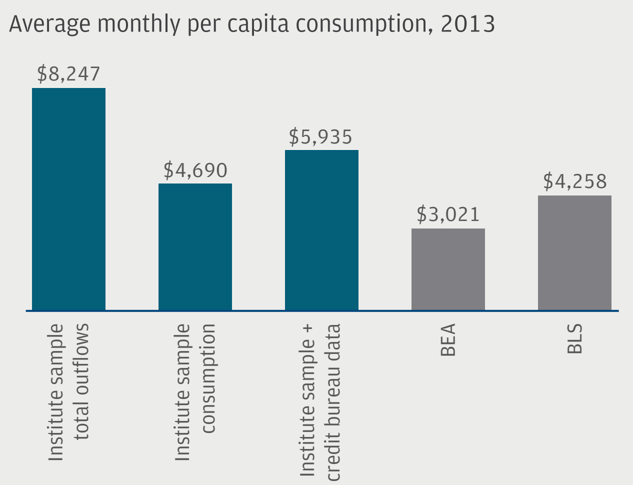
FIGURE 18: INCOME CHARACTERISTICS OF THE INSTITUTE SAMPLE COMPARED TO NATIONAL BENCHMARKS



* Census estimates of percentage receiving any labor income and percentage receiving Social Security are based on the population aged 25 and older.

Source: Bureau of Economic Analysis data are from the National Income and Product Accounts, 2013 and represents total disposable income. Census data come from the Current Population Survey and are based on person income estimates from 2013.

FIGURE 19: CONSUMPTION CHARACTERISTICS OF THE INSTITUTE SAMPLE COMPARED TO NATIONAL BENCHMARKS



Source: Bureau of Economic Analysis data are from the National Income and Product Accounts, 2013 and reflects per capita consumption. Bureau of Labor Statistics data come from the Current Expenditure Survey and are based on average annual expenditures.

Future Enhancements

Our new and evolving data set, mined at regular intervals for changing trends, provides fresh insights into the volatility of income and consumption that comprise this report. Additionally, throughout this section we highlighted a few new facts that demonstrate how the bank's consumer data can provide fresh insights into the financial lives and behaviors of individuals and the economy at large. Interesting in their own right, these facts also highlight the granularity and power of our data asset.

The JPMorgan Chase Institute will continue to build and refine this data asset to address an even broader array of important economic and policy questions pertaining to individuals and households. Ultimately, our ability to understand where individuals spend their money and how this varies month to month is an important cornerstone of our data asset. This inaugural report focuses on the volatility observed in income and consumption and how income and consumption changes move together. In future reports, we plan to provide additional insights into the financial ins and outs observed in our customer data. For example, if income declines, we will see if individuals cut back on restaurant purchases and increase what they spend on groceries. In addition, we plan to widen the time horizon of the data by including a full five-year history and including real-time, new monthly data as they occur. This will allow us to more fully explore the impacts of the global financial crisis and produce timely indices that can benchmark financial behavior. Other planned expansions to the data asset include a more complete view of consumer assets and liabilities to develop a perspective on household balance sheets. Finally, while still fully preserving the anonymity of our data, we plan to add third-party data on demographics to develop a granular perspective on consumer finance issues by important segments of the population and household characteristics.

Unique JPMorgan Chase Assets

While our inaugural report and initial data investment focus entirely on consumer finance, the future research agenda of the JPMorgan Chase Institute extends across the portfolio of JPMorgan Chase's lines of business and vast geographic reach. Future data assets and analytics of the JPMorgan Chase Institute will focus on businesses, large and small, the global flows of funds and other critical economic topics. These data, combined with expert insights, are unique assets the JPMorgan Chase Institute will use to provide a comprehensive perspective on the complex inner workings of the global economy and help policymakers, businesses and nonprofit leaders make smarter decisions to advance global prosperity.

Glossary

Channel: The delivery channel by which money flows in or out of an account. Outflow channels include debit card purchase, ACH – debit, check withdrawal and ATM cash withdrawal. Inflow channels include ACH – credit, ATM cash deposit, ATM check deposit and teller deposit.

Consumption: Outflow transactions that have been identified by the JPMorgan Chase Institute as spending. These include purchases of goods and services, utilities, tax payments, ATM withdrawals, debt payments, rent, non-principal portion of mortgage payments and fees. Transfers to other financial institutions, allocations to saving and investment accounts and outflow transactions that cannot clearly be identified as consumption are not included in consumption.

Consumption mobility: The degree to which individuals move between consumption quintiles from one year to the next, where consumption quintiles are defined based on the distribution of consumption in the current year.

Credit bureau data: Monthly data obtained from credit bureaus on all lines of credit a de-identified individual has, as reported by financial institutions including JPMorgan Chase & Co.

Credit utilization: The size of the individual's revolving balance across all open credit cards expressed as a percentage of the total credit limit across all open credit cards. The revolving credit card balance is estimated as the total outstanding credit card balance minus the credit card spending in that month.

Income: Inflow transactions that have been identified by the JPMorgan Chase Institute as income. These include direct deposits such as payroll, annuities and dividends, tax refunds, unemployment insurance, Social Security and ATM deposits. Transfers from other financial institutions, saving and investment accounts, and inflow transactions that cannot clearly be identified as income are not included in income.

Income mobility: The degree to which individuals move between income quintiles from one year to the next, where income quintiles are defined based on the distribution of income in the current year.

Income quintile: One of the five segments of the population where each segment reflects 20% of the population on the basis of the income distribution. Quintile 1 refers to individuals with incomes in the bottom 20% in terms of income (0–20%); quintile 2 refers to individuals in the 20%–40% range of incomes; quintile 3 refers to individuals in the 40%–60% range of incomes; quintile 4 refers to individuals in the 60%–80% range of incomes; and quintile 5 refers to individuals in top 20% in terms of income (80%–100%).

Inflow: A credit transaction to an account holder's checking account.

JPMorgan Chase data asset: The evergreen data set compiled by the JPMorgan Chase Institute that currently includes monthly balances on all Chase consumer accounts and credit bureau data on liabilities for 2.5 million primary account holders as well as daily transaction-level data on Chase debit and credit cards for a random sample of 100,000 account holders.

Liquid asset: Cash and assets readily accessible at no or minimal cost, including balances held in checking, savings and money market deposit accounts and money market funds.

Outflow: A debit transaction to an account holder's checking account.

Primary account holder: The signatory legally responsible for the account. In the JPMorgan Chase data asset, all account activity is reflected under the person listed as the primary account holder. When there is more than one primary account holder, the account activity is reflected under the person listed first on the account.

Responders: Individuals for whom income and consumption changes fell within 10 percentage points of each other between 2013 and 2014. Examples include those who saw between 2013 and 2014 a 10% increase in income and a 15% increase in consumption, or a 10% decrease in income and a 15% decrease in consumption.

Sticky Optimists: Individuals for whom consumption changes positively exceed income changes by at least 10 percentage points between 2013 and 2014. Examples include those who saw between 2013 and 2014 a 10% increase in income and a 21% increase in consumption, or a 21% decrease in income and a 10% decrease in consumption.

Sticky Pessimists: Individuals for whom income changes positively exceed consumption changes by at least 10 percentage points between 2013 and 2014. Examples include those who saw between 2013 and 2014 a 21% increase in income and a 10% increase in consumption, or a 10% decrease in income and a 21% decrease in consumption.

Transaction: A single deposit or withdrawal of funds by any transaction channel.

Volatility: The magnitude of positive and negative dispersions from the median.

References

Bennett, B., Conover, D., O'Brien, S., & Advincula, R. (2014). Cash Continues to Play a Key Role in Consumer Spending: Evidence from the Diary of Consumer Payment Choice. *Federal Reserve Bank of San Francisco, FedNotes, April.*

Boshara, R., & Emmons, W. (2012). After the fall: rebuilding family balance sheets, rebuilding the economy. *Federal Reserve Bank of St. Louis Annual Report,* 4–15.

Browning, M., Crossley, T., & Winter, J. (2014). The measurement of household consumption expenditures. Institute of Fiscal Studies Working Papers (W14/07).

Campos, R., Reggio, I., & García-Píriz, D. (2012). Micro vs. macro consumption data: the cyclical properties of the consumer expenditure survey. Economics Working Papers we1220, Universidad Carlos III, Departamento de Economía.

Chetty, R., Hendren, N., Kline, P., Saez, E., & Turner, N. (2014). Is the United States still a land of opportunity? Recent trends in intergenerational mobility. *National Bureau of Economic Research,* no. w19844.

Dahl, M., DeLeire, T., & Schwabish, J. (2008). Recent trends in the variability of individual earnings and household income. *Washington, DC: US Congressional Budget Office.*

DeBacker, J. M., Heim, B. T., Panousi, V., & Vidangos, I. (2012). Rising Inequality: Transitory or Permanent? New Evidence from a Panel of US Tax Returns 1987–2006. *Indiana University-Bloomington: School of Public & Environmental Affairs Research Paper Series,* (2011-01), 01.

Dynan, K., Elmendorf, D., & Sichel, D. (2012). The evolution of household income volatility. *The BE Journal of Economic Analysis & Policy,* 12(2).

Federal Deposit Insurance Company. (2014). 2013 FDIC National Survey of Unbanked and Underbanked Households. *Federal Deposit Insurance Company.*

Fisher, J. D., & Johnson, D. S. (2006). Consumption mobility in the United States: Evidence from two panel data sets. *Topics in Economic Analysis & Policy,* 6(1).

Gorbachev, O. (2011). Did household consumption become more volatile? *The American Economic Review,* 2248-2270.

Gosselin, P. & Zimmerman, S. (2008). Trends in Income Volatility and Risk, 1970-2004. *Urban Institute Working Paper, May.*

Gottschalk, P., & Moffitt, R. (2009). The rising instability of US earnings. *The Journal of Economic Perspectives,* 3–24.

Hall, R. (1978). Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence. *Journal of Political Economy,* 86(6).

Hardy, B., & Ziliak, J. P. (2012). Decomposing Trends in Income Volatility: The 'Wild Ride' at the Top and Bottom. *University of Kentucky Center for Poverty Research Discussion Paper Series*, DP2012-02.

Hastings, J. S., & Washington, E. L. (2008). The First of the Month Effect: Consumer Behavior and Store Responses. *National Bureau of Economic Research,* no. w14578.

Kaplan, G., Violante, G. L., & Weidner, J. (2014). The Wealthy Hand-to-Mouth. *Brookings Papers on Economic Activity (Spring).*

Keys, B. J. (2008). Trends in *Income and Consumption Volatility. In Income Volatility and Food Assistance in the United States,* Dean Jolliffe, James P. Ziliak, editors. Kalamazoo, MI: W.E. Upjohn Institute, pp. 11–34.

Maestas, N., Mullen, K. J., & Strand, A. (2013). Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt. *The American Economic Review,* 103(5).

Morduch, J., & Schneider, R. (2013). Spikes and Dips: How Income Uncertainty Affects Households. *US Financial Diaries Issue Brief. New York: Financial Access Initiative at New York University.*

Mullainathan, S., & Shafir, E. (2013). *Scarcity: Why having too little means so much.* Macmillan.

Parker, J. A. (2014). Why Don't Households Smooth Consumption? Evidence from a 25 million dollar experiment. *Manuscript, July.*

Stephens, M. (2003). "3rd of the Month": Do Social Security Recipients Smooth Consumption Between Checks? *American Economic Review,* 406–422.

Stephens, M. (2006). Paycheck Receipt and the Timing of Consumption. *The Economic Journal,* 6(7).

The Pew Charitable Trusts. (2015a). The Precarious State of Family Balance Sheets. January.

The Pew Charitable Trusts. (2015b). Americans' Financial Security. March.

Endnotes

1 In each case here and throughout much of this report, we have calculated symmetric percent change between A and B, calculated as (B-A)/(0.5*(A+B)). This formula has the benefit of allowing for positive and negative changes to be represented symmetrically and also for changes from zero to be calculable.

2 During this period, total inflows observed decreased by 2.2% and outflows increased by 0.2% respectively between 2013 and 2014.

3 See U.S. Economy at a Glance: Perspective from the BEA accounts, available at http://www.bea.gov/newsreleases/glance.htm. In addition, increases in income tax rates that took effect on January 1, 2013 caused some employers to pay out 2012 bonuses in 2012 rather than 2013, and thus potentially resulted in lower incomes in 2013.

4 In our sample 12% experienced at least a 25% decline in income between 2013 and 2014, and 31% experienced a change in income of 25% or more in either direction. Dynan et al (2012), using the Panel Study of Income Dynamics, found that the percentage of people experiencing a 25% or more decline in income over a two-year period increased from 16% in the early 1970s to over 20% in the 2000s. A 2008 Congressional Budget Office study found that roughly 20% of the population experienced a 25% decline in income between 2002 and 2003, and 39% experienced earnings changes of more than 25% in either direction between 2001 and 2002. In terms of aggregate income mobility, our data confer with previous estimates of income mobility. For example, DeBacker et al (2012) find that 74% of individuals remained in the same income quintile from year to year between 1987 and 2009 compared to 72% in our sample.

5 Our finding that consumption is significantly more volatile than income sharply departs from theoretical predictions from the permanent income hypothesis from economics that people should be able to smooth consumption as they experience transitory income shocks and only adjust consumption in response to permanent changes in income (Hall, 1978). It is also inconsistent with existing empirical research, which shows volatility of food consumption to be significantly lower than volatility of income. Gorbachev (2011) and Keys (2008), using the PSID, found that year-over-year food consumption volatility is substantially lower than income volatility and that volatility in food consumption grew at less than half the rate that income volatility grew between 1970 and the early 2000s. Fisher and Johnson (2006) complemented the PSID with data from the Consumer Expenditure Survey in order to estimate both income mobility and consumption mobility in the United States and found them to be similar.

6 Here income and consumption quintiles are created based on the JPMorgan Chase Institute's estimates of income and consumption respectively. In 2013 income quintiles are defined as follows: quintile 1 is $16,200 or less; quintile 2 is $16,200–$28,900; quintile 3 is $28,900–$43,200; quintile 4 is $43,200–$67,600; and quintile 5 is $67,600 and above. In 2013 the consumption quintile 1 is $29,800 or less; quintile 2 is $29,800–$43,400; quintile 3 is $43,400–$61,000; quintile 4 is $61,000–$92,600; and quintile 5 is $92,600 and above.

7 A slightly smaller sample with an odd number of individuals was used for this calculation, resulting in approximate quintiles that cause the shares moving up or down a quintile to not be exactly equal.

8 Data presented in this report have not been seasonally adjusted.

9 See The JPMorgan Chase Institute Data Asset section for a full discussion of our transaction classification strategy.

10 For the purposes of this analysis, we base income quintiles on an annual pre-tax income estimate for 2014 ascertained by JPMorgan Chase based on individual, third-party and zip code-level data rather than the income estimated by the JPMorgan Chase Institute analysis of inflows. The first income quintile is $35,300 or less; quintile 2 is $35,300–$50,000; quintile 3 is $50,000–$67,800, quintile 4 is $67,800–$100,000; and quintile 5 is $100,000 or more. As in earlier analyses, we continue to use symmetric percent change.

11 For the sake of comparison, we calculate the distribution of percentage changes in income between four-month periods over a 16-month period between 2010 and 2011 using data from the Survey of Income and Program Participation (SIPP) for both the entire national sample and a subsample with monthly income always $400 or greater in order to approximate our sample selection screen. We discover that for the entire national sample income volatility in the bottom income quintile far exceeds income volatility in the top income quintile. Income volatility is overall much lower in the subsample of individuals surveyed in the SIPP with income greater than $400, and there is comparable income volatility across income quintiles.

12 We do not believe that we have underestimated volatility of consumption in the bottom quintile to the same extent that we may have underestimated volatility of income in the bottom quintile. Our sampling approach requires only that people have five outflow transactions rather than any minimum dollar amount.

13 We performed two robustness checks to validate these results. First, we calculated the 25th and 75th percentile changes for total inflows and outflows in order to ensure that our results are not driven by irregularities in the way in which we categorized inflows and outflows into income and consumption respectively or biases in how volatile uncategorized flows (e.g., paper checks) are relative to categorized flows. We find that volatility is even greater when we evaluate total inflows and comparable when we evaluate total outflows and volatility increases with income quintile. For example the 25th to 75th percentile spread on inflows was -22% to 24% for income quintile 1 and -32% to 33% for income quintile 5, wider spreads than those in Figure 7. The 25th to 75th percentile spread on outflows was -24% to 25% for income quintile 1 and -30% to 31% for income quintile 5, comparable to the spreads in Figure 7. The second robustness check was to calculate the 25th to 75th spreads on a small sub sample of roughly 8,000 people for whom 90% of total inflow and outflow dollars were fully categorized. We also segment this group into income quintiles based on the income identified by the JPMorgan Chase Institute rather than by the annual income estimate. We find that income volatility among this subsample is virtually identical to the results presented in Figure 7, but that consumption volatility is slightly lower than what we observed above (e.g., -22% to 24% for middle income quintile earners). Bottom quintile earners experience slightly less volatility in income than top quintile earners, still likely due to our sampling criteria that select out people with inflows below $500, but they experience slightly more consumption volatility (-24% to 26% spread) than individuals in the top income quintile (-21% to 23% spread).

14 For example, Gosselin and Zimmerman (2008) showed that income volatility was not only higher among bottom quintile earners than among top quintile earners and but also increased significantly more between 1973 and 2003 than volatility for top quintile earners. Hardy and Ziliak (2012) showed that volatility of earnings among the top 1% of earners has been increasing but still remains lower than the volatility experienced by the bottom 10%. The recent evidence from the U.S. Financial Diaries highlights the extent and unpredictability of fluctuations in income experienced by low-income families and the impact on their ability to cover costs (Morduch and Schneider, 2013). Even recent research on the negative impacts of the Great Recession largely concentrates on the economically vulnerable subgroups (Boshara and Emmons, 2012).

15 Each dot in Figures 8 and 9 represents a group of individuals in order to adhere to privacy protocols.

16 Although there is a large literature that explores the relationship between income and consumption changes, it typically explores the short-term impacts of income on spending and demonstrates the significant immediate increase in spending in response to positive income fluctuations such as the 2008 Economic Stimulus Payment (Parker, 2014), Social Security benefits (Stephens, 2003), food stamp benefits (Hastings and Washington, 2008) and even paychecks (Stephens, 2006).

17 Specifically, we measure credit utilization by estimating the total revolving balance (i.e., outstanding balance that individuals carry from the previous month) as a percentage of the total credit limit across all credit cards.

18 Refer to the section on the JPMorgan Chase Institute Data Asset for a more complete discussion of how our sample differs from then nation.

19 The 5th percentile change in income from the prior month by income quintile was -76% for quintile 1, -81% for quintile 2, -83% for quintile 3, -90% for quintile 4 and -101% for quintile 5. The 95th percentile change in consumption from the prior month by income quintile was 80% for quintile 1, 81% for quintiles 2 and 3, 82% for quintile 4 and 87% for quintile 5. Pre-tax median income as reported by the Survey of Consumer Finance was $14,203 for quintile 1, $28,407 for quintile 2, $46,668 for quintile 3, $76,090 for quintile 4, and $121,744 for the 80th to 90th percentiles within quintile 5. We calculated post-tax median incomes by assuming tax rates of 15% for quintile 1; 25% for quintiles 2, 3 and 4; and 28% for quintile 5.

20 We recognize that some individuals may intuitively consider their median or mean levels of income and consumption as more relevant reference points than the previous month when experiencing and managing volatility. Although we believe measuring monthly volatility as the change from the previous month is more indicative of the liquidity management challenge; as a robustness check, we also calculated liquid asset buffers using the percentage changes in income and consumption relative to the moving average and moving median levels over the prior 12 months. These methodologies yielded slightly lower estimates of the liquid asset buffer necessary for individuals in each income quintile to weather volatility: $1,200 for quintile 1; $2,100–$2,200 for quintile 2; $3,600 for quintile 3; $6,000 for quintile 4; and $9,800–$10,700 for quintile 5. On the other hand, as noted previously, our estimates of income volatility, and therefore liquid asset buffers, may likely be biased downward given that our sampling criteria require individuals to have a minimum of $500 in deposits each month. We intend to continue to refine these estimates as we further explore these methodological and sampling approaches.

21 A recent study by Brookings describes roughly a third of the population as the "wealthy hand-to-mouth," because, although they have illiquid assets, they do not have sufficient liquid assets to cover cash flow needs or other unexpected shocks (see Kaplan et al, 2014). Similarly, a recent Pew study highlights that even the middle class do not have sufficient resources to weather the financial fluctuations they experience (The Pew Charitable Trusts, 2015a).

22 See several studies by Sendhil Mullainathan and Eldar Shafir and others recently summarized in their book: *Scarcity: Why Having too Little Means So Much*.

23 Several countries, such as Mexico, South Africa and the United Kingdom, have established banking infrastructures that support general-purpose immediate fund transfers.

24 Response rates to these surveys are typically in the range of 60% to 90%, but have been decreasing in recent years; according to Browning et al (2014), CEX response rates fell 11 percentage points from 1986 to 2007.

25 See Morduch and Schneider (2013) and The Pew Charitable Trusts (2015a and 2015b).

26 See for example, Chetty et al (2014) and Maestas et al (2013).

27 Among our sample, roughly half of primary account holders are individual account holders, and the activity we see for these individuals is likely to reflect the financial life of one person. The other half of our sample are primary account holders on at least one individual account, but who also have a joint account. The account activity we see for these individuals could reflect the financial lives of multiple individuals if they are the primary account holder on the joint account, or it could offer only a partial view of their financial life if they are the secondary account holder on the joint account.

28 In fact our sample includes individuals in all 50 states.

29 16% of our sample has an unidentified gender. We have displayed the gender distribution of those with an identified gender.

30 The Census Bureau's estimate of monthly household income for 2013 was $6,053 before taxes. Although the primary account holder is our unit of analysis, some accounts may reflect the financial lives of more than one individual.

31 For example, if individuals are paying their credit card out of their Chase account this consumption will be reflected in our total outflow numbers, but we may not have identified it as consumption per se if the individual pays their credit card bill by writing a paper check. If, however, individuals use some other non-Chase financial account to pay these credit card bills, or they spend using but don't pay off non-Chase credit cards, this activity will not be reflected in the total outflows we observe.

32 The Consumer Expenditure Survey measures average annual consumption per consumer unit, which essentially includes all members of a household and reflects the consumption of, on average, 2.5 people. With an average annual consumption per consumption unit of $51,100, the average per person average annual consumption is $20,440, or $1,703 on a monthly basis, which is significantly lower than the Personal Consumption Expenditure as measured by the Bureau of Economic Analysis. For a discussion and explanation of these discrepancies, see Campos et al (2012).

Zillow® Real Estate and Rental Data: Why We're Different

Introduction

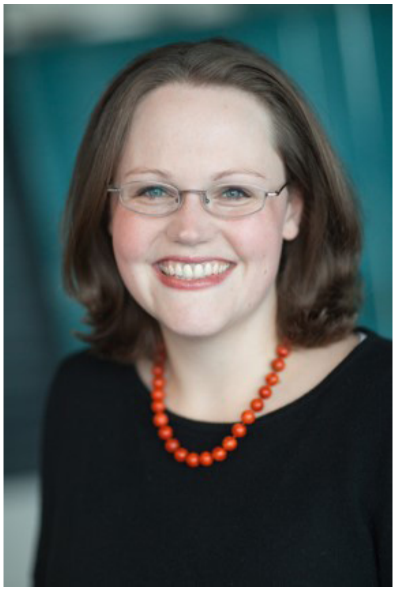
Zillow® is the leading real estate and home-related information marketplace. Zillow is dedicated to empowering consumers with data, inspiration and knowledge around the place they call home.

Zillow operates an industry-leading economics and analytics bureau led by Zillow's Chief Economist, Dr. Stan Humphries. At Zillow, Dr. Humphries and his team of economists and data analysts produce extensive housing data and analysis covering more than 500 markets nationwide. Zillow Research produces various real estate, rental and mortgage-related metrics and publishes unique analyses on current topics and trends affecting the housing market.

At Zillow's core is our living database of more than 100 million U.S. homes, featuring both public and user-generated information including number of bedrooms and bathrooms, tax assessments, home sales and listing data of homes for sale and for rent. This data allows us to calculate, among other indicators, the Zestimate, a highly accurate, automated, estimated value of almost every home in the country as well as the Zillow Home Value Index and Zillow Rent Index, leading measures of median home values and rents.

About Zillow Research

Zillow Research is a division of the analytics group of Zillow, dedicated to analysis and research into various topics pertaining to the housing market. Zillow researches various topics involving housing, including home values, rents, mortgage and rent affordability, negative equity, mortgages and more. Zillow data has been used by leading industry, government and academic institutions. Zillow is committed to data transparency, and all of our data can be downloaded directly from www.zillow.com/data.



Dr. Svenja Gudell, Chief Economist

Svenja joined the company in 2011 and leads the industry-leading economic research team at Zillow, a recognized voice of impartial, data-driven economic analysis on the U.S. housing market. Under Svenja's leadership, Zillow produces monthly reports on housing trends for more than 450 metros nationwide, down to the ZIP code level. In addition, Svenja and her team publish original research on various real estate topics, ranging from rental and mortgage affordability, negative equity and forecasting, to policy, generational and mortgage research. Svenja has presented to various federal agencies and at numerous industry conferences, and has been widely quoted in national and local media.

Prior to joining Zillow, Svenja did economic, financial and strategy consulting for Analysis Group and was an Assistant Economist in the Research Group of the Federal Reserve Bank of New York. Svenja has a Bachelor of Arts in Economics from the University of Rochester, a Master of Arts in Economics from New York University, a Master of Science in Business Administration from the University of Rochester, and a Ph.D. in Finance from the University of Rochester.

Major Reports

Real Estate Market Reports:
Released monthly three weeks after the close of the period. Offers forecasts and includes data on home values, rents and forecasts for the U.S. and hundreds of local markets.

Negative Equity:
Released quarterly. Includes data on underwater homeowners, delinquency rates and loan-to-value distributions.

Affordability Indices:
Released quarterly. Data covers price-to-income ratios and the share of income spent on typical rent or mortgage payments for a median-priced home leased or purchased during the quarter.

Breakeven Horzizon:
Released quarterly. Analyzes the number of years a consumer needs to live in a home to make owning that home more financially advantageous than renting the same home.

Case-Shiller Forecast:
Released hours after the prior month's S&P Case-Shiller Home Price Index release. Includes a prediction of the Case-Shiller Home Price Index for the 10-City, 20-City and National Indices.

Zillow Home Price Expectations Survey:
Released quarterly. Home price forecasts and opinions on housing trends from a topical survey of more than 100 leading economists and housing experts.

Zillow Housing Confidence Index:
Released twice per year. Measures consumer confidence in local housing markets, both currently and looking forward, including their attitudes towards homeownership and their homeownership aspirations.

Real Estate and Rental Data

Real Estate Market Reports

The Zillow Real Estate Market Reports offer an overview of national and local real estate and rental markets. The reports are compiled by Zillow Real Estate Research and are released monthly and quarterly approximately three weeks after the close of the period. The reports include forecasts, foreclosure data, for-sale inventory, home values and rents.

The following metrics and more are produced regularly as a part of the Zillow Real Estate Market Reports and are available for free download at www.zillow.com/data. Most data is available at the national, state, metro (CBSA), county, city, ZIP code and neighborhood levels.

- List prices
- Sale prices
- Home sales
- Home value per square foot
- List price per square foot
- Sale-to-list price ratio
- Percent of listings with price cuts
- Median percentage of price cuts
- Percent of homes sold for loss/gain
- Homes foreclosed
- Foreclosure re-sales
- Percent of homes increasing/decreasing in value
- Percent of homes sold in the past year (turnover)
- Price-to-rent ratio
- Rental listing prices

Below are detailed descriptions of our most popular indices and metrics:

Zillow Home Value Index (ZHVI)

The Zillow Home Value Index is the median home value for a given area. It is reported monthly and includes single-family, condominium and cooperative homes. Unlike median sale price, ZHVI measures the value of all homes, regardless of whether the homes sold within a given month. It is expressed in dollars and is seasonally adjusted. ZHVI is published where available at the national, state, metro (CBSA), county, city, ZIP code and neighborhood levels.

Why use the Zillow Home Value Index?

Repeat sales indices like the Case-Shiller Home Price Index measure the difference in sale prices of the same home over time – regardless of whether one of the sales was a foreclosure re-sale. Distressed sales are often well below market value, and when they make up a larger portion of the market – as they have in recent years – can skew overall values wildly. ZHVI includes non-distressed sales information and estimated non-distressed home values in an effort to avoid these swings. The ZHVI has an extensive footprint, covering 95 percent of the U.S. housing stock by market value. This broad coverage includes the same large, coastal markets included in leading indices, and complements that data with more comprehensive information from the larger number of smaller, less volatile markets nationwide. Reports like those produced by the National Association of Realtors® or Redfin® measure only the median sale price of homes sold in a reporting period. This leads to a bias toward what has sold in a particular period. ZHVI includes data on virtually all homes in a given area, regardless of whether they have recently sold or not, resulting in a more accurate representation of the true median value of an area's housing stock.

Zillow Rent Index (ZRI)

The Zillow Rent Index is the median estimated monthly rental price for a given area, and covers single-family, condominium, and cooperative homes in Zillow's database, regardless of whether they are currently listed for rent. It is expressed in dollars and is seasonally adjusted. The Zillow Rent Index is published where available at the national, state, metro (CBSA), county, city, ZIP code and neighborhood levels.

Zillow Home Value Forecast (ZHVF)

The Zillow Home Value Forecast uses past home value trends and data on current market conditions, including home sales, months of housing inventory supply and unemployment, to predict home value performance over the next 12 months for the country and for more than 250 local markets nationwide. It is published where available at the national, state, metro (CBSA), county, city, ZIP code and neighborhood levels.

Negative Equity

The Zillow Negative Equity Report calculates the share and number of homeowners in an area who are underwater on their mortgage, owing more than the value of their home. The report includes, but is not limited to, negative equity rates, "effective" negative equity rates (homeowners with less than 20 percent equity in their home), loan-to-value ratios and delinquency rates. To calculate negative equity, the estimated value of a home is matched to TransUnion data on all outstanding mortgage debt and lines of credit associated with the home, including home equity lines of credit and home equity loans. All personally identifying information is removed by TransUnion. The Zillow Negative Equity Report uses actual, outstanding mortgage debt balances, obtained from TransUnion, in our calculations. Competitors like CoreLogic® estimate outstanding mortgage debts, a less precise methodology. Overall, this report covers more than 870 metros, 2,400 counties, 17,500 cities and 23,000 ZIP codes nationwide.

Real Estate and Rental Data

Affordability

Zillow's affordability indices measure the share of income needed to purchase or rent a typical home nationwide and in hundreds of metro areas. Produced quarterly, our affordability analysis compares the share of income needed to rent or buy currently, over time and compared to long-term, historical averages. For-sale affordability assumes a buyer making an area's median household income and purchasing the median-valued home, with 20 percent down and at prevailing interest rates. Rent affordability assumes a renter making an area's median household income and renting the median-priced rental property. Affordability metrics are published both as standalone analyses, and in combination with homeownership rates, employment rates, measures of confidence and more to determine how affordability – or lack thereof – is impacting local housing markets.

For-Sale Inventory

Zillow Research produces raw and smoothed, seasonally adjusted time series of the for-sale inventory of homes. Each week, the number of single-family, condominium and cooperative housing units listed for sale on Zillow are counted. This listing count includes standard, real estate owned (REO) and for-sale by owner (FSBO) listings, but excludes pending, auction and new construction, as well as Zillow's Make Me Move and Coming Soon listings. The median of these counts within a month is calculated as the monthly level of inventory, and a seasonal adjustment is applied. This seasonally adjusted series is then smoothed using a three-month rolling average. Inventory is available at the national, state, metro (CBSA), county, city and ZIP code levels.

Buyer-Seller Index

The Zillow Buyer-Seller Index combines the sale-to-list price ratio, the percent of homes subject to a price cut and the time properties spend on the market. In this analysis, a sellers' market is not necessarily one where home values are rising, but rather one in which homes are on the market for a shorter time, price cuts occur less frequently and homes are sold at prices very close to (or greater than) their last list price. In buyers' markets, homes for sale stay on the market longer, price cuts occur more frequently and homes are sold for less relative to their listing price. The Buyer-Seller Index is available at the metro area level for comparisons across metro areas, and at the city and ZIP code level for comparison within a metro area.

Breakeven Horizon (Rent vs. Buy)

Zillow's breakeven horizon calculates the point, in years, at which buying a home becomes less expensive than renting the same home. It incorporates all costs associated with buying and renting, including upfront payments, closing costs, anticipated monthly rent and mortgage payments, insurance, taxes, utilities and maintenance costs. It also factors in historic and anticipated home value appreciation rates, rental prices and rental appreciation rates. The Breakeven Horizon is available at the national, state, metropolitan area (CBSA), county, city, ZIP code and neighborhood levels.

Zillow Housing Confidence Index

The Zillow Housing Confidence Index (ZHCI) measures consumer confidence in local housing markets, both currently and over time. As a forward-looking indicator, the ZHCI can also help determine future trends in consumer spending and overall housing demand. It is based on a unique, national survey that collects more than 300,000 responses from more than 10,000 Americans. The overall ZHCI is made up of three separate components, which measure consumers' perceptions of current market performance, their expectations for future performance and their attitudes toward the value of housing in general and its social impact. The ZHCI is calculated twice annually for the U.S. as a whole and for 20 of the largest metro areas, covering more than 100 million Americans. The ZHCI is sponsored by Zillow and conducted by Pulsenomics LLC.

Zillow Home Price Expectations Survey

The quarterly Zillow Home Price Expectations Survey is sponsored by Zillow and conducted by Pulsenomics LLC. The survey asks more than 100 leading economists, real estate experts and investment and market strategists to predict the path of the U.S. Zillow Home Value Index over the next five years as well as a short topical survey on current issues affecting the housing market.

Zillow Mortgage Access Index

The Zillow Mortgage Access Index measures seven distinct credit and lending variables to determine how easy or difficult it is to obtain a mortgage, both currently and over time. The monthly index starts in 2002 (Jan. 2002 = 100), just before the housing bubble really started inflating. Index values above 100 indicate it is easier to get a mortgage than it was in 2002, and values less than 100 indicate it is more difficult. The seven variables included in the index are: Credit scores of successful applicants; the debt-to-income ratio of successful applicants; the share of low-down-payment loans that are privately insured; the prevalence of second mortgages; the level of non-conforming loans; the spread between 30-year, fixed-rate mortgages and 10-year treasury rates; and the number of quotes given to lower and higher-credit borrowers on Zillow Mortgages.

Case-Shiller Forecast

Zillow Real Estate Research forecasts the S&P/Case-Shiller national, 10- and 20-City Home Price Indices one month before the numbers are officially released. A few hours after Case-Shiller releases new data, Zillow forecasts the next month's data, with a median absolute error of 0.2 percent across all forecasts.

Real Estate and Rental Data

Days on Zillow

We estimate the median days on market of homes sold in a given month. For a single observation, we must have a date on an official transaction record and a listing record on Zillow for that same address. We smooth the series with a simple three-month symmetrical moving average, which weights the center observation twice as heavily as the days on market observed in adjacent months. This series is available for select counties, metro (CBSA) and states.

Market Health Index

Zillow calculates the Market Health Index on a scale from 0 to 10, with 0 being the unhealthiest and 10 being the healthiest, illustrating the current health of a region's housing market relative to other markets across the country. The Market Health Index is formed from ten different metrics, including: Monthly change in ZHVI, annual change in ZHVI, percent of homes selling for a gain, the Zillow Home Value Forecast, Days on Zillow, the number of foreclosed homes, foreclosure re-sales, negative equity, delinquency rate and unsold REOs. The Market Health Index is available at the city, county, metro (CBSA), state and ZIP code level.

Home Sales

We produce a time series of new and existing home sales – arms-length transactions of single-family, condominium and cooperative homes on the national, metro and county levels. The series dates to June 2008, and the transaction date is the closing date recorded on the county deed. The home sales time series are adjusted for latency in county reporting of home sales. Data is available for selected counties and metro areas.

New and Existing Home Sales Forecast

We forecast both existing home sales data from the National Association of Realtors and new home sales data from the U.S. Census Bureau, using both historical data and a model of housing market fundamentals. Our models seek to explain the economic factors driving home sales each month. We release a forecast of the reports a few days before the scheduled data release.

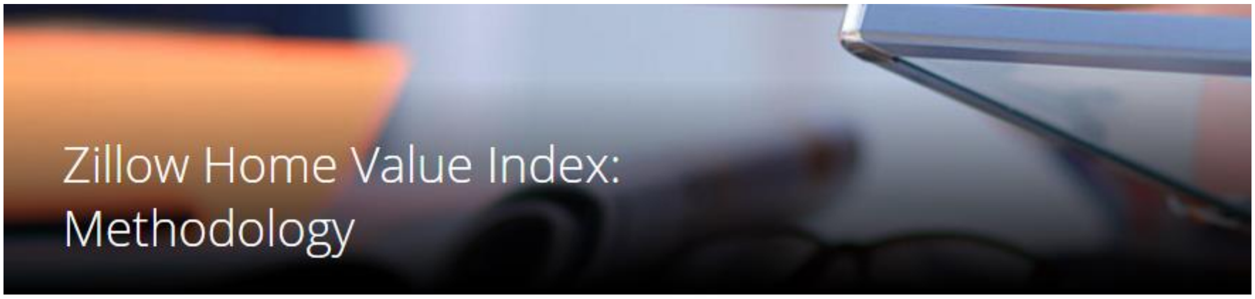
Quick Comparison of Housing Indices and Data

Zillow	• The Zillow Home Value Index (ZHVI) tracks the median value of all homes in an area, regardless of whether the home was sold during the reporting period. • ZHVI includes only arms-length transactions and does not include distressed sales. ZHVI covers single-family residences, condominiums and co-op homes and is available monthly for metro areas, states, cities, counties and ZIP codes. ZHVI is also available for only single-family residences. • Median list and sales prices are also available from Zillow.
S&P/Case-Shiller	• The S&P Case-Shiller (SPCS) Home Prices Indices track repeat sales of pre-existing single-family homes in 20 metropolitan areas. SPCS does not include condominiums or co-op homes and homes must have sold at least twice to be included in the indices. • The SPCS Indices include foreclosures, which bias the change in home prices upwards or downwards depending on market conditions.
National Association of Realtors® (NAR)	• NAR computes median sales prices of existing homes, released quarterly by metro area which is based off a monthly survey of its members. Data is available for the nation, four regions and approximately 170 metro areas and covers single family residences, condominiums and co-ops beginning in 2005. NAR estimates that it covers only 30 to 40 percent of all existing home sales with its survey. • Median list and sales prices are biased measures of the true value of homes, as they only include data on homes that were listed for sale or sold during the reporting period.
Redfin	• Redfin reports median sales prices and median list prices of only MLS-listed homes for the 53 areas in which it operates. Does not include for-sale by owner or off-market homes. • Median list and sales prices are biased measures of the true value of homes, as they only include data on homes that were listed for sale or sold during the reporting period.

About Zillow

Zillow® is the leading real estate and rental marketplace dedicated to empowering consumers with data, inspiration and knowledge around the place they call home, and connecting them with the best local professionals who can help. In addition, Zillow operates an industry-leading economics and analytics bureau led by Zillow's Chief Economist Dr. Svenja Gudell. Dr. Gudell and her team of economists and data analysts produce extensive housing data and research covering more than 450 markets at Zillow Real Estate Research. Zillow also sponsors the quarterly Zillow Home Price Expectations Survey, which asks more than 100 leading economists, real estate experts and investment and market strategists to predict the path of the Zillow Home Value Index over the next five years. Zillow also sponsors the bi-annual Zillow Housing Confidence Index (ZHCI) which measures consumer confidence in local housing markets, both currently and over time. Launched in 2006, Zillow is owned and operated by Zillow Group (NASDAQ: Z), and headquartered in Seattle.

Contact: researchfeedback@zillow.com



Zillow Home Value Index: Methodology

Home / Methodology

By Andrew Bruce on 1/3/2014 Share

Introduction

In setting out to create a new home price index, a major problem Zillow sought to overcome in existing indices was their inability to deal with the changing composition of properties sold in one time period versus another time period. Both a median sale price index and a repeat sales index are vulnerable to such biases (see the analysis here for an example of how influential the bias can be). For example, if expensive homes sell at a disproportionately higher rate than less expensive homes in one time period, a median sale price index will characterize this market as experiencing price appreciation relative to the prior period of time even if the true value of homes is unchanged between the two periods.

The ideal home price index would be based off sale prices for the same set of homes in each time period so there was never an issue of the sales mix being different across periods. This approach of using a constant basket of goods is widely used, common examples being a commodity price index and a consumer price index. Unfortunately, unlike commodities and consumer goods, for which we can observe prices in all time periods, we can't observe prices on the same set of homes in all time periods because not all homes are sold in every time period.

The innovation that Zillow developed in 2005 was a way of approximating this ideal home price index by leveraging the valuations Zillow creates on all homes (called Zestimates). Instead of actual sale prices on every home, the index is created from estimated sale prices on every home. While there is some estimation error associated with each estimated sale price (which we report here), this error is just as likely to be above the actual sale price of a home as below (in statistical terms, this is referred to as minimal systematic error). Because of this fact, the distribution of actual sale prices for homes sold in a given time period looks very similar to the distribution of estimated sale prices for this same set of homes. But, importantly, Zillow has estimated sale prices not just for the homes that sold, but for all homes even if they didn't sell in that time period. From this data, a comprehensive and robust benchmark of home value trends can be computed which is immune to the changing mix of properties that sell in different periods of time (see Dorsey *et al.* (2010) for another recent discussion of this approach).

For an in-depth comparison of the Zillow Home Value Index to the Case Shiller Home Price Index, please refer to the Zillow Home Value Index Comparison to Case-Shiller

Each Zillow Home Value Index (ZHVI) is a time series tracking the monthly median home value in a particular geographical region. In general, each ZHVI time series begins in April 1996. We generate the ZHVI at seven geographic levels: neighborhood, ZIP code, city, congressional district, county, metropolitan area, state and the nation.

Underlying Data

Estimated sale prices (Zestimates) are computed based on proprietary statistical and machine learning models. These models begin the estimation process by subdividing all of the homes in United States into micro-regions, or subsets of homes either near one another or similar in physical attributes to one another. Within each micro-region, the models observe recent sale transactions and learn the relative contribution of various home attributes in predicting the sale price. These home attributes include physical facts about the home and land, prior sale transactions, tax assessment information and geographic location. Based on the patterns learned, these models can then estimate sale prices on homes that have not yet sold.

The sale transactions from which the models learn patterns include all full-value, arms-length sales that are not foreclosure resales. The purpose of the Zestimate is to give consumers an indication of the fair value of a home under the assumption that it is sold as a conventional, non-foreclosure sale. Similarly, the purpose of the Zillow Home Value Index is to give consumers insight into the home value trends for homes that are not being sold out of foreclosure status. Zillow research indicates that homes sold as foreclosures have typical discounts relative to non-foreclosure sales of between 20 and 40 percent, depending on the foreclosure saturation of the market. This is not to say that the Zestimate is not influenced by foreclosure resales. Zestimates are, in fact, influenced by foreclosure sales, but the pathway of this influence is through the downward pressure foreclosure sales put on non-foreclosure sale prices. It is the price signal observed in the latter that we are attempting to measure and, in turn, predict with the Zestimate.

Market Segments

Within each region, we calculate the ZHVI for various subsets of homes (or market segments) so as to afford greater insight into what is happening in a particular market. All market segments are shown in the table below. Only residential properties are included in the ZHVI calculation. Non-residential properties, such as office buildings, shopping centers and farms are not included.

One very useful form of market segmentation that we produce is based on the distribution of home values within the metropolitan area. Here we assign properties into one of three tiers based on their Zestimates on a particular date: top, middle or bottom tier. The thresholds for the price tiers vary from metro to metro and are determined by the distribution of home values in each metro. Since Zestimates are time-dependent, a property may belong to different price tiers at different dates. To reduce tier switching, we exclude properties near the boundaries of price tiers when assigning tiers. Thus, the sum of Zestimates in all three tiers does not equal the number of Zestimates for the “All Homes” market segment.

Table 1: Market Segments for Zillow Home Value Index

Market Segment	Number of Zestimates	Description
All Homes	87.3 M	Single family + condominium + cooperative
Single Family	78.1 M	Single family only
Condo	9.2 M	Condominium + cooperative only
0 or missing	31.6 M	0 Bedroom or missing
1 Bedroom	1.7 M	1 Bedroom
2 Bedroom	11.1 M	2 Bedroom
3 Bedroom	28.6 M	3 Bedroom
4 Bedroom	11.7 M	4 Bedroom
5+Bedroom	2.7 M	5 Bedroom or more
Top Tier	27.0 M	Top price tier among homes within the same metropolitan
Middle Tier	27.0 M	Middle price tier among homes within the same metropolitan
Bottom Tier	27.0 M	Bottom price tier among homes within the same metropolitan

Methodology

Using the estimated market value of every home as represented in the Zestimate, the main steps in the construction of the ZHVI are as follows:

1. Calculate Raw Median Zestimates
2. Adjust for Any Residual Systematic Error
3. Apply Henderson Moving Average Filter
4. Apply Seasonal Adjustment
5. Final Quality Control

Calculating Raw Median Zestimates

Let t be a discrete independent time variable with a value at the end of each month. Let $\mathbf{H}(t)$ be an M by N matrix with each element $h_{ij}(t)$ representing the number of homes at time t for the i-th market segment in the j-th geographical region, where M is the total number of market segments and N is the total number of unique regions having a minimum required number of Zestimates. Currently, we have M=12 and N=77,590. Geographical regions include national, state, metro, county, city, ZIP code and neighborhood. The Number of Zestimates column in Table 1 above represents the number of homes in the i-th element of h_{ij} when j='National' and t='Nov-2013'.

Let $\boldsymbol{z}_{ij}(t)$ be the vector of Zestimates of all homes at time t having length $h_{ij}(t)$ for i-th market segment and j-th region. The raw median Zestimate, $r_{ij}(t)$, for i-th market segment and j-th region is defined as:

$r_{ij}(t)=\text{Median}(\boldsymbol{z}_{ij}(t))$

$r_{ij}(t)$ is the median Zestimate and is an element of the M by N matrix $\mathbf{R}(t)$. In order to ensure reliability and stability, we only compute r_{ij} when $h_{ij}(t)$ is above some minimum threshold. For November 2013, there are a total of 389,451 market segments by regions for which the median could be computed:

Count$\{r_{ij}(t) \neq \text{NA}$, for i=1,..M and j=1,..N$\}$ is 389,451.

Table 2: Number of regions by market segment having raw median Zestimates

Market Segment	National	State	MSA	County	City	Neighborhood	Zip	Total
All Homes	1	51	917	2,830	23,057	8,664	24,460	59,980
Single Family	1	51	917	2,828	22,976	8,068	24,249	59,090
Condo	1	51	507	895	4,189	2,916	6,629	15,188
0 or missing	1	51	868	2,464	14,023	3,304	15,097	35,808
1 Bedroom	1	51	537	1,097	2,112	1,080	3,418	8,296
2 Bedroom	1	51	742	1,821	9,173	3,083	11,870	27,461
3 Bedroom	1	51	817	2,105	13,310	5,523	15,796	37,603
4 Bedroom	1	51	766	1,829	8,633	3,124	11,485	25,889
5+Bedroom	1	51	619	1,249	3,524	1,018	5,648	12,110
Top Tier	1	51	913	1,681	12,554	4,112	14,862	34,184
Middle Tier	1	51	913	1,704	14,058	4,877	16,364	37,968
Bottom Tier	1	51	913	1,676	12,941	5,119	15,173	35,874

Adjust for Any Residual Systematic Error

Zestimate errors are both time and region dependent. While the errors produced by the Zestimate algorithm are generally equally distributed above and below the actual sale price, there can be some residual systematic error detected once more historical sales are known (systematic error here is defined as the median raw error being slightly greater or less than zero). In this event, raw median Zestimates are adjusted through the use of a correction factor in the manner described below.

Let $u_{ij}(t)$ be the median home value free of systematic error. Then, the raw median Zestimate can be expressed in terms of $u_{ij}(t)$ as:

$r_{ij}(t) = \{1 + b_j(t)\} * u_{ij}(t)$

where $b_j(t)$ is the systematic error in Zestimates representing the median fluctuation of Zestimates above or below the actual sold prices within the time window centered around t for the j-th region. We calculate the Zestimate systematic error as:

$b_j(t) = \text{Median}(\{\mathbf{z}_j(t-1) - \mathbf{s}_j(t)\}/\mathbf{s}_j(t))$

where $\mathbf{s}_j(t)$ is a vector of sale prices and $\mathbf{z}_j(t-1)$ are Zestimates corresponding to the same properties as $\mathbf{s}_j(t)$ but with the estimated sale price taken from the period immediately prior to the actual sale (to ensure that the estimate has not been influenced by the sale). The vector of sales, $\mathbf{s}_j(t)$, is obtained through the following approach:

1. Find all sales within a 30-day window centered on t.
2. Increase the window on either side of time t until at least 100 transactions are obtained for region j.
3. The maximum length of the window is 365 days.
4. For time t at the two endpoints of the time series, a maximum window length of 182 days is imposed.
5. Fit a natural cubic smoothing spline with knots evenly spaced every twelve months to the time series $b_j(t)$ to remove noise.
6. If fewer than 100 transactions are present, then shrink the $b_j(t)$ towards zero.

After computing $b_j(t)$, the adjusted median of Zestimates is an M by N matrix $\mathbf{U}(t)$ calculated as:

$u_{ij}(t) = r_{ij}(t)/\{1 + b_j(t)\}$

Apply Henderson Moving Average Filter

We apply a 5-term Henderson moving average filter to $\mathbf{U}(t)$ to reduce noise in the data. This filter was derived by Henderson, R. (1916). The filter weights applied in the middle of a time series are symmetric, while the end filter weights are asymmetric.

$\mathbf{MA}(t) = w_1\mathbf{U}(t-2) + w_2\mathbf{U}(t-1) + w_3\mathbf{U}(t) + w_4\mathbf{U}(t+1) + w_5\mathbf{U}(t+2)$

where

w = (-0.07343, 0.29371, 0.55944, 0.29371, -0.07343) for the middle points: t = 3, 4, .. , TMax-2

w = (-0.04419, 0.29121, 0.52522, 0.22776, 0) for t = TMax -1

w = (0, 0.22776, 0.52522, 0.29121,-0.04419) for t = 2

w = (-0.13181, 0.36713, 0.76467,0, 0) for the end point: t = TMax

w = (0, 0, 0.76467, 0.36713, -0.13181) for the start point: t = 1

The resultant M by N matrix $\mathbf{MA}(t)$ is a smooth estimate of the median home value free of residual systematic error. This may not be as necessary for large regions such as the nation and states because of the large available data set, but it is applied to all levels for consistency.

Apply Seasonal Adjustment

Home sales are affected by seasons within the same year. Adjusting for seasonality is desirable so that the trend is more apparent for ease of comparison and forecasting. Since Zestimates and the ZHVI depend on sale prices, the time series **MA**(t) does contain some seasonality. We remove this seasonality using a seasonal-trend decomposition procedure (STL) based on the Loess method developed by Cleveland *et al*. (1990). STL is a filtering procedure for decomposing a time series into seasonal, trend and remainder components:

$$\mathbf{MA}(t) = \mathbf{S}(t) + \mathbf{T}(t) + \mathbf{RE}(t)$$

where **S**(t), **T**(t) and **RE**(t) are the seasonal, trend and remainder components respectively. We remove seasonality by adding the trend and remainder components to form the seasonally adjusted ZHVI:

$$\mathbf{ZHVI}(t) = \mathbf{T}(t) + \mathbf{RE}(t)$$

The remainder component, **RE**(t), represents irregular features in the time series which we preserved.

Final Quality Control

The time series matrix **ZHVI**(t) has the same dimension as **H**(t) which is M by N (as noted, 12 x 77,590). While this theoretically could produce almost 1 million different time series, in practice many time series are eliminated because of data sparseness or temporal volatility. A four-star quality rating function is applied to every ZHVI time series. The variables feed to this function are features associated with each ZHVI time series. They include:

1. Number of Zestimates
2. Number of transactions in most recent three months
3. Temporal volatility measured by annualized, monthly and quarterly change
4. Number of outliers
5. Gaps
6. Jumps
7. Disclosure or non-disclosure states

After suppressing those with star ratings below 2, there are 242,362 unique deliverable ZHVI time series for the report period ending November 2013.

Table 3: Number of deliverable ZHVI time series by region level and market segment

Market Segment	National	State	MSA	County	City	Neighborhood	Zip	Total
All Homes	1	49	485	947	10,647	5,176	12,602	29,907
Single Family	1	49	486	946	10,623	4,955	12,507	29,567
Condo	1	48	347	600	3,467	2,087	5,434	11,984
0 or missing	1	49	425	813	6,072	2,116	7,315	16,791
1 Bedroom	1	47	311	539	1,469	762	2,388	5,517
2 Bedroom	1	48	447	840	6,376	2,911	8,649	19,272
3 Bedroom	1	48	511	998	8,746	4,140	11,131	25,575
4 Bedroom	1	50	509	956	6,863	2,708	9,346	20,433
5+Bedroom	1	46	426	769	3,073	959	4,986	10,260
Top Tier	1	47	533	915	8,744	3,030	10,899	24,169
Middle Tier	1	50	493	860	9,943	3,814	11,676	26,387
Bottom Tier	1	48	486	831	7,927	3,302	9,905	22,500

Restatement of the ZHVI

ZHVIs for all geographic regions and market segments are updated every month. Since there is variable latency in Zillow's receipt of transactional data from public records, Zillow's estimate of residual systematic error can change as new transactions arrive. Historical estimates of systematic error are recalculated monthly and incorporated into revised ZHVI time series. As a result, there can be restatements of the ZHVI for up to three years from initial reporting date.

Impact of Methodology Change: November 2013

With the release of November 2013 Zillow Home Value Index (ZHVI) data, we have improved the underlying valuation model, introduced additional data filtering algorithms and developed a new approach to dealing with residual systematic error. The result of these changes led to a 24.1% increase in the number of regions for which Zillow reports a ZHVI (see Table 4). The historical values for the ZHVI have been restated with these changes, leading to slightly higher current estimate of the median home value nationally. The revised ZHVIs are qualitatively similar to the ZHVI computed using the previous methodology, although the new time series are significantly smoother.

Table 4: Increase in reporting regions by region type from the previous (Oct. 2013) and new (Nov. 2013) methodologies

	Oct. ZHVI	**Nov. ZHVI**	**Increase**
States	44	49	11.4%
Metros	389	485	24.7%
Counties	744	947	27.3%
Cities	8,535	10,647	24.7%
Neighborhoods	4,190	5,176	23.5%
ZIP Codes	10,205	12,602	23.5%
Total	**24,107**	**29,906**	**24.1%**

The new valuation model and data filtering algorithms have led to a restatement of past values for the ZHVI. Figures 1 and 2 compare the ZHVI for the old versus new methodology for the US and for the Composite 20 metropolitan markets. The revision of the ZHVI has generally raised the overall estimated median value of homes. The national median home value is higher by 3.2%: $168,000 versus $162,800 with the previous methodology. The increase is due to better accuracy of the new valuation model and better screening of transactions that are normally excluded from the ZHVI (e.g., foreclosures and foreclosure re-sales).

Figure 1: Comparison of new and old ZHVI methodologies (U.S)

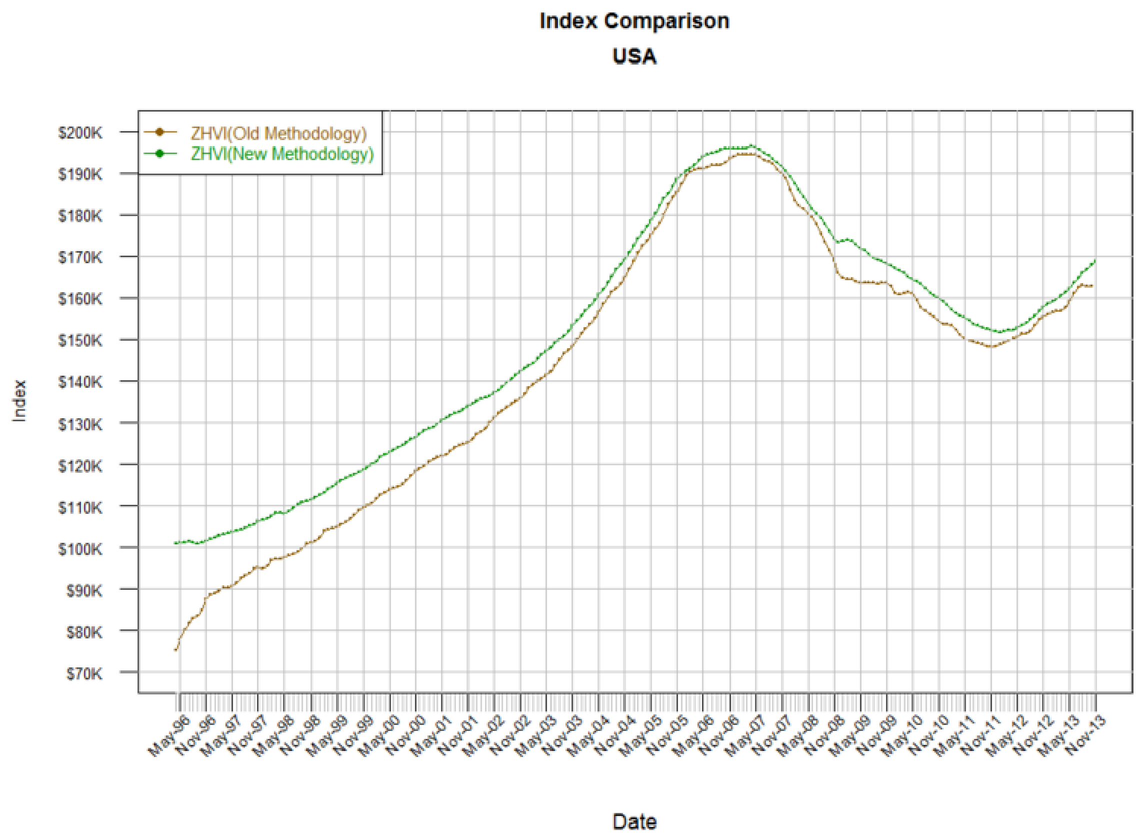
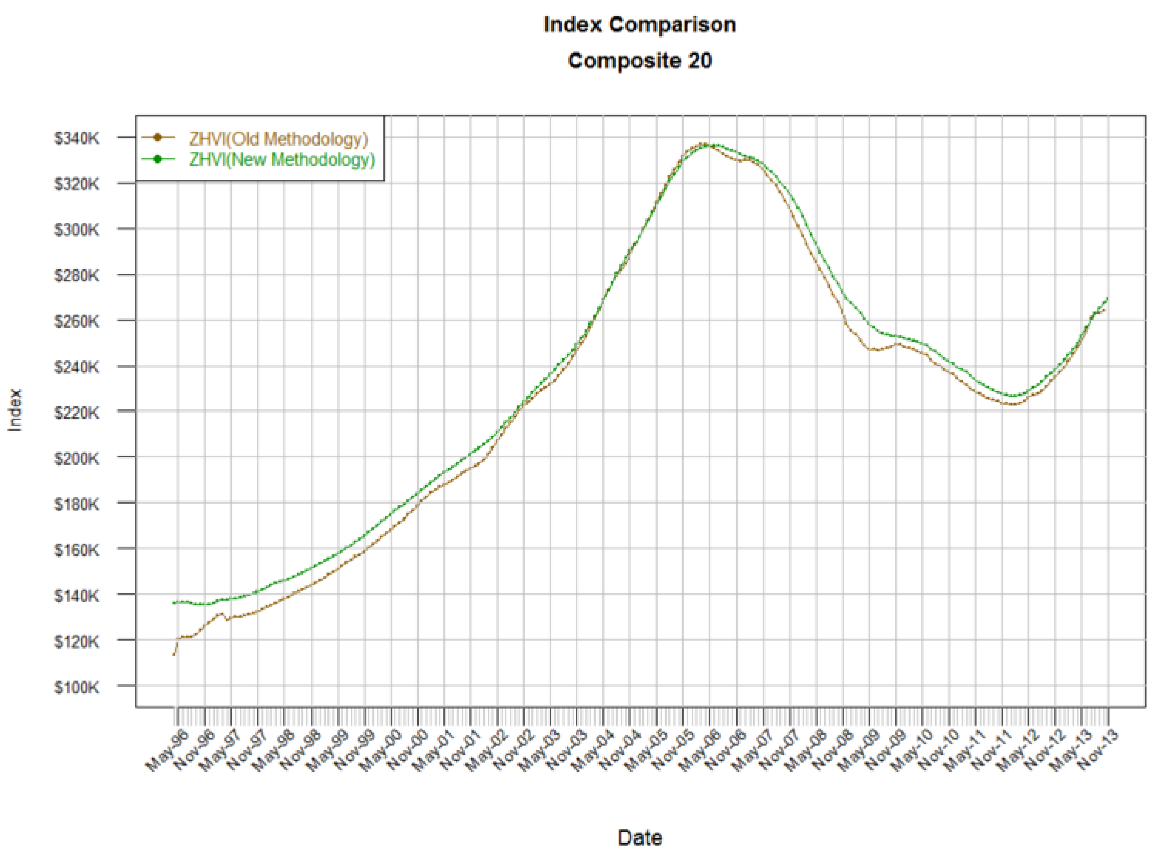


Figure 2: Comparison of new and old ZHVI methodologies (Composite 20)



Qualitatively, the restated ZHVIs are similar to the ZHVIs calculated using the previous methodology, although they have less volatility, particularly on a shorter time scales. The size and direction of the revision depends very much on the region, although the ZHVI year-over-year change is typically revised downward for areas that have experienced high price appreciation. For example, the November year-over-year change for Phoenix has been revised downward from 19.4% to 15.0%. In addition to revisions due to the more accurate valuation method and improved transaction filtering, the new approach to correcting residual systematic error also contributes somewhat to the restatement of the index levels.

November 2013 Methodology Revision Details

The methodology improvements released with the November 2013 data were based on three main areas. First, the ZHVI has been rebased on the latest valuation model that produces the Zestimates. Second, new and improved filtering algorithms were incorporated to screen out bad transactional data. Third, the approach to estimate the systematic error correction was updated. These are discussed below.

More Accurate Valuation Model: The November 2013 ZHVI release incorporated the latest version of the home valuation model, which is the model that produces the Zestimate. This latest version resulted in a significant increase in the accuracy of the Zestimate (which is 13% more accurate than a year ago). Accuracy was especially improved for high-end homes (30% improvement), waterfront homes and homes in less urban areas. The new valuation model resulted in moderate revisions to the national ZHVI, resulting in a small increase the overall level of the ZHVI and somewhat damped peak-to-trough cycles.

Improved Transaction Filtering: The November 2013 ZHVI also took advantage of improved filtering on transactional data. This change impacts the ZHVI indirectly through a corresponding improvement in the valuation model (see above) and more directly through more accurate correction for residual systematic error. The valuation algorithm and the ZHVI exclude transactions that are not representative of what is considered a full-value, arms-length transaction between a buyer and a seller. This definition excludes transactions such as foreclosures, foreclosure re-sales, estate sales and intra-family transfers. In doing a better job of identifying these transactions, the ZHVI has increased on a national basis as well as in many regions. For example, the current ZHVI for Sacramento is $300,000 under the revised methodology versus $284,500 under the previous methodology, a 5.4% increase in level.

Systematic Error Correction: The systematic error correction is based on comparing the transactions versus the Zestimate for a time period. Since transactions are relatively sparse, particularly in smaller geographic regions, the new systematic error correction method smooths the bias over time and shrinks the estimate towards zero. The smoothing procedure is based on fitting a natural cubic spline with knots evenly spaced every twelve months. Specifically, the smoothed value is given by the predicted value from the model.

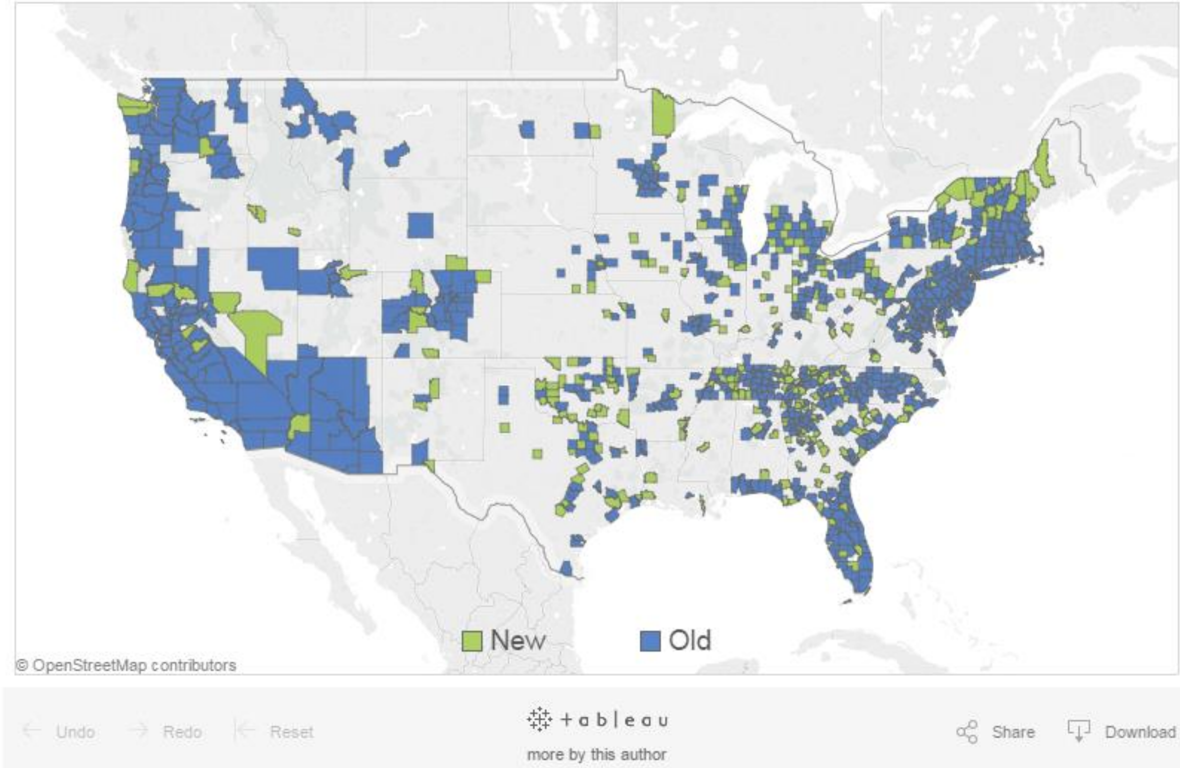
$$y_{i,t} = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \sum_{k=1,\cdots,K} b_k (t - \xi_k)^3$$

For regions with fewer than 100 transactions in a time period, the resulting smoothed estimate of bias will be shrunk towards zero.

New Coverage of ZHVI by County

New coverage of ZHVI (green) in addition to the old coverage (blue) is shown in the interactive map below:

ZHVI Coverage by County



References

Henderson, R. (1916). Note on Graduation by Adjusted Average. Transactions of the American Society of Actuaries, 17, 43-48.

Cleveland, R.B, Cleveland, W.S., McRae, J.E., and Terpenning, I. (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics, 6*, 3–73.

Dorsey, R.E., Hu, H., Mayer, W.J., & Wang, H. (2010). Hedonic versus repeat-sales housing price indexes for measuring the recent boom-bust cycle. *Journal of Housing Economics, 19* (2), 75-93.

Zillow® Home Value Index

The Zillow Home Value Indices (ZHVI) are the most accurate and timely measures of residential real estate prices in the United States. The indices are available for more than 350 metropolitan statistical areas representing more than 95% of the total housing stock by value. The index family includes breakdowns by different geographic units, from neighborhoods to counties to the entire country. The ZHVI are also available for home type, price tier and number of bedrooms. The indices are available more than a month before other indices for the same reporting period, making the ZHVI the leading indicator of housing prices.

Living Database of Homes

At the core of the ZHVI is Zillow's proprietary "living database" of more than 110 million homes. The database integrates information from disparate sources, including prior sales, county records, tax assessments, real estate listings, mortgage information and GIS data. As a consumer-facing company, our database is enriched by homeowners who have claimed their homes on Zillow and edited home facts. Data on more than a third of homes in our database have been updated by users, giving Zillow a much better picture of the housing stock than is provided by public records alone.

From Data to Zestimate®

For more than 100 million homes, Zillow calculates a Zestimate, an estimate of value for each home. The Zestimate is based on a suite of sophisticated "automated valuation models" (AVM). The models are re-trained three times a week based on the latest data, and each home's Zestimate is updated daily. Zillow is the only firm to publish both the accuracy of its AVM's and the entire valuation history for each home. The Zestimate is unbiased for each region and price tier, meaning just as many Zestimates are likely to be higher than the actual value of the home as they are likely to be below the value.

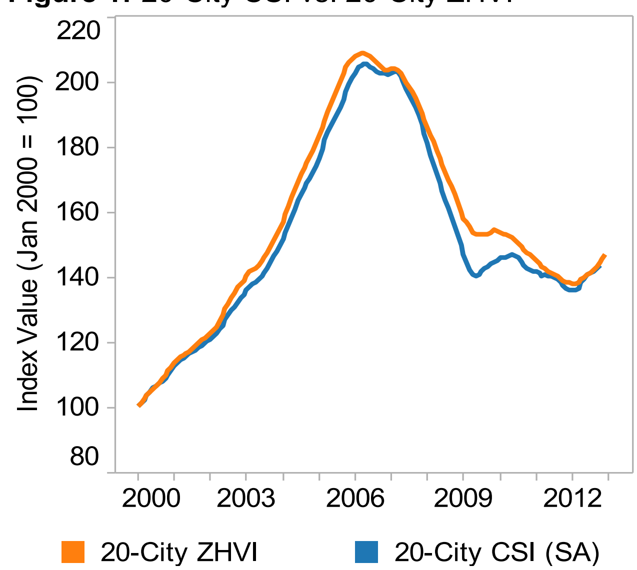
From Zestimate to ZVHI

The ZHVI is defined as the median of all Zestimates in a region or price tier. The beauty of the ZHVI is its simplicity: It is straightforward and intuitive. By leveraging the Zestimate, the ZHVI confers several benefits, including timeliness, accuracy and most importantly, lack of bias. Since the ZHVI is an estimate based on all homes, it doesn't require convoluted models to try to correct for biases inherent in other approaches, such as repeat sales methodology. The ZHVI is published monthly roughly three weeks after the end of the reporting period.

Advantages Over Case-Shiller

Case-Shiller is based on repeat sales methodology, which measures price change by collecting data on homes that have resold in a given region. Case-Shiller only includes homes that have sold at least twice in recent history and excludes all new construction. Because segments of homes may appreciate at different rates and those segments are not proportionally represented in the mix of repeat sales, the index may be biased. The bias is especially acute at smaller geographic regions where limited repeat sales data is available. The table below provides a comprehensive overview of the differences between the Case-Shiller family of indices and the ZHVI.

Figure 1: 20-City CSI vs. 20-City ZHVI



Exclusion of Foreclosure Re-Sales

Foreclosure resales are substantially different from non-distressed sales, and are often priced at a substantial discount to the non-distressed value of a home. Because of this fact, foreclosure resales are not used to train the AVM's underlying the Zestimate. This means that the ZHVI excludes foreclosures from the price index. By contrast, Case-Shiller includes foreclosure resales in their indices, leading them to represent a blend of two very different market segments. Consumers and investors interested in understanding the change in home values in a regular, non-distressed market will get a less accurate estimate by looking at Case-Shiller.

Figure 1 shows the impact foreclosure resales have on the Case-Shiller indices. The 20-City Composite Case-Shiller Home Price Index (CSI) compares quite well to the 20-city composite created from the ZHVI for most of the historical period. The two indices diverge beginning in 2008 as the number of foreclosure resales begins to increase, and they converge again in 2011 as the discount associated with foreclosure resales diminishes (in 2012, the median national discount of a foreclosure resale relative to a non-foreclosure sale was only 7.4%).

Differences in Footprint

When looking at the national level, we find substantial differences between the CSI and ZHVI (see Figure 2) because the data footprint of the CSI is smaller than that of the ZHVI. The more inclusive ZHVI shows a less dramatic boom in home prices in the 2001 to 2006 period (since most of the less populated areas of the country did not experience a housing boom) and a commensurately smaller decline during the bust. As noted, the decline in home prices is further exaggerated in the CSI by the inclusion of foreclosure resales. Interestingly, the peak of home values in the ZHVI was mid-2007 versus 2006 for CSI, again a better representation of the full country versus only the large coastal metros that experienced home price bubbles. Further evidence that the National CSI is biased toward the large coastal metros that make up the 20-City CSI is the strong similarity of these two indices as shown in Figure 3.

In short, while there is fundamental methodological difference between the ZHVI and CSI, most actual differences in the two indices historically are attributable to the differing treatment of foreclosure resales and the difference in the data footprint.

Forecast Case-Shiller Indices from ZHVI

Because the ZHVI is published over a month before the Case-Shiller index, Zillow is able to forecast the value of Case-Shiller before it is released by adjusting for the mix of foreclosures in a region. Since May 2011, the median error for the Zillow forecasts of the Case-Shiller indices has been 0.1% for the 20-City Composite and 0.2% for the 10-City Composite.

Figure 2: National CSI vs. National ZHVI

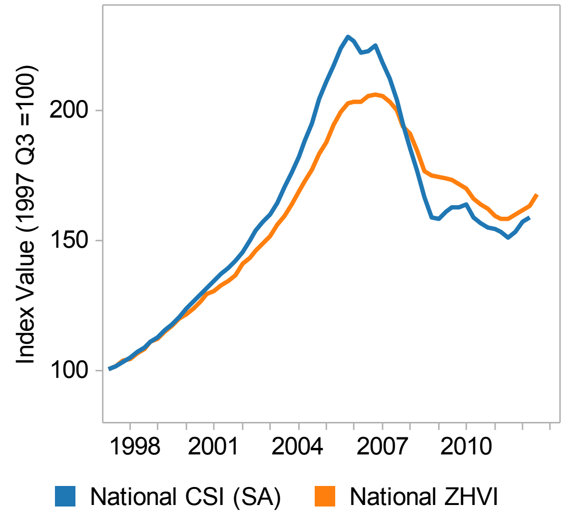
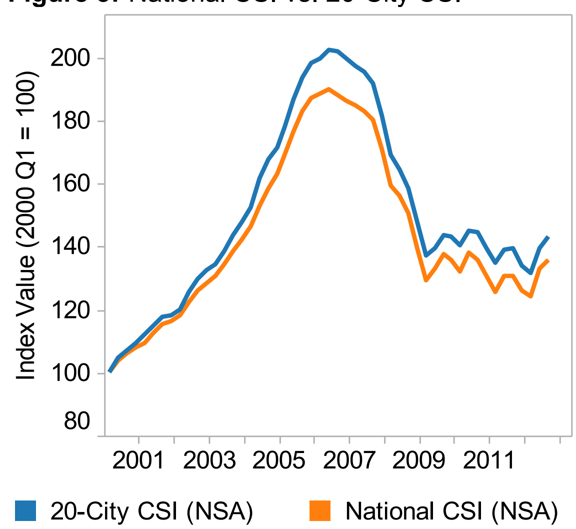


Figure 3: National CSI vs. 20-City CSI



Summary Comparison of National Home Price Indices

	S&P/Case-Shiller© U.S. National Home Price Index ("S&P/CSI")	**Zillow National Home Value Index ("ZHVI")**
Primary Purpose	Benchmark for home price-linked financial product development, trading and settlement	Housing market analysis and research
Methodology	Repeat Sales · Weighted composite of 9 underlying Census Division repeat sales indices · Seasonally- & non-seasonally-adjusted versions	Hedonic Imputation · Median of actual and estimated market values of all homes within a market (or market segment) · 3-month smoothed, using a Henderson Filter · Seasonally-adjusted only
Underlying Data	Sale pairs for single-family homes only, i.e., SF homes for which: · A sale price (distressed or non-distressed) is recorded within the current index reporting period and for which a prior historical sale price is also available · Excludes newly-constructed homes · Index data history to 1987	Actual and estimated values of 83 million individual single-family homes, condos, and co-ops: · Actual, non-distressed sale prices recorded during the index reporting period · Estimated non-distressed market values for every home in the Zillow database that does not sell during the reporting period · Includes newly constructed homes · Index data history to 1997
Coverage	@71% of US housing stock by market value	@95% of US housing stock by market value
Release Frequency	Quarterly	Monthly
Reporting Lag	56 – 61 days	18-23 days

Zillow Rent Index: Methodology

By Yeng Bun on 3/12/2012

Introduction

Similar to the Zillow Home Value Index (ZHVI), we created the Zillow Rent Index (ZRI) to track the monthly median rent in particular geographical regions. Like the ZHVI, we sought to create an index for rents that is unaffected by the mix of homes for rent at any particular time. This makes temporal comparisons of rents more valid since the index is tracking the rents for a consistent stock of inventory. It also makes it easier to compare the ZHVI and ZRI since they are based on a similar set of homes whereas traditional metrics tracking rent and sale prices are often based on markedly different sets of homes (often located in different neighborhoods), thus making comparisons less valid.

Underlying Data

Similar to the Zestimate, we estimate rents (Rent Zestimates) based on proprietary statistical and machine learning models. Within each county or state, the models observe recent rental listings and learn the relative contribution of various home attributes in predicting prevailing rents. These home attributes include physical facts about the home, prior sale transactions, tax assessment information and geographic location as well as the estimated market value of the home (Zestimate). Based on the patterns learned, these models estimate rental prices on all homes, including those not presently for rent. The purpose of the Rent Zestimate is to give consumers an indication of the fair market rent for a home, while the purpose of the ZRI is to give consumers insight into rental price trends in a way that is not biased by the mix of homes currently for rent.

Because of the availability of Zillow rental listing data used to train the models, Rent Zestimates are only available back to November 2010 and, consequently, each ZRI time series begins on this date as well. We generate the ZRI at seven geographic levels including neighborhood, ZIP code, city, congressional district, county, metropolitan area, state and the nation.

Market Segments

Within each region, we calculate the ZRI for various subsets of homes (or market segments) so as to afford greater insight into what is happening in a particular market. All market segments are shown in the table below. Apartments are treated as condominiums for training purposes only and aren't included in the final index. For more details about market segments, please see the Zillow Home Value Index methodology.

Table 1: Market Segments for Zillow Rent Index

Market Segment	Number of Rent Zestimates	Description
All Homes	84.9 M	Single family + condominium + cooperative
Single Family	75.5 M	Single family only
Condo	09.3 M	Condominium + cooperative only
0 or missing	42.9 M	0 Bedroom
1 Bedroom	01.7 M	1 Bedroom
2 Bedroom	11.4 M	2 Bedroom
3 Bedroom	29.4 M	3 Bedroom
4 Bedroom	12.2 M	4 Bedroom
5+Bedroom	03.3 M	5 Bedroom or more
Top Tier	26.2 M	Top price tier among homes within the same metropolitan
Middle Tier	26.2 M	Middle price tier among homes within the same metropolitan
Bottom Tier	26.2 M	Bottom price tier among homes within the same metropolitan

Methodology

Using the estimated rent of every home as represented in the Rent Zestimate, the main steps in the construction of the ZRI are as follows:

1. Calculate Raw Median Rent Zestimates
2. Apply Simple 3-Month Moving Average
3. Final Quality Control

Calculate Raw Median Rent Zestimates

Let t be a discrete independent time variable with a value at the end of each month. Let $\mathbf{H}(t)$ be an M by N matrix with each element $h_{ij}(t)$ representing the number of homes at time t for the i-th market segment in the j-th geographical region, where M is the total number of market segments and N is the total number of unique regions having a minimum required number of Rent Zestimates. Currently, we have M=12 and N=57,022. Geographical regions include national, state, metro, county, city, ZIP code, neighborhood and congressional district. The Number of Rent Zestimates column in Table 1 above represents the number of homes in the i-th element of h_{ij} when j='National' and t='Jan-2012'.

Let $\boldsymbol{z}_{ij}(t)$ be the vector of Rent Zestimates of all homes at time t having length $h_{ij}(t)$ for i-th market segment and j-th region. The raw median Rent Zestimate, $r_{ij}(t)$, for i-th market segment and j-th region is defined as:

$r_{ij}(t)$=Median($\boldsymbol{z}_{ij}(t)$)

$r_{ij}(t)$ is the median Rent Zestimate and is an element of the M by N matrix $\mathbf{R}(t)$. In order to ensure reliability and stability, we only compute r_{ij} when $h_{ij}(t)$ is above some minimum threshold. For Jan 2012, there are a total of 391,375 unique set of regions and market segments for which the median could be computed:

Count{$r_{ij}(t) \neq$NA, for i=1,..M and j=1,..N} is 391,375.

Table 2 shows the counts of Rent Zestimates by region level and market segment. For example, we have usable data to calculate raw medians in up to 2,485 counties for the single-family home market segment.

Market Segment	National	State	MSA	County	Congressional District	City	Neighborhood	Zip
All Homes	1	51	848	2,486	433	21,229	8,475	22,672
Single Family	1	51	848	2,485	433	21,154	7,810	22,482
Condo	1	51	460	806	410	4,032	3,035	6,467
0 or missing	1	51	825	2,301	433	17,078	4,772	18,894
1 Bedroom	1	51	488	956	416	2,305	1,115	3,757
2 Bedroom	1	51	713	1,680	432	9,331	3,899	12,040
3 Bedroom	1	51	784	1,964	433	13,130	5,497	15,594
4 Bedroom	1	51	739	1,689	432	8,819	3,443	11,676
5+Bedroom	1	51	586	1,160	430	4,123	1,672	6,673
Top Tier	1	49	838	1,523	429	11,892	3,984	14,070
Middle Tier	1	49	838	1,549	429	13,469	4,773	15,712
Bottom Tier	1	49	839	1,511	428	12,069	5,186	14,374
Total	**12**	**606**	**8,806**	**20,110**	**5,138**	**138,631**	**53,661**	**164,411**

Apply Simple Three-Month Moving Average

We apply a simple three-month moving average to **R**(*t*) to filter out noise in the data:

ZRI(*t*)={ **R**(*t*)+**R**(*t-1*)+ **R**(*t-2*)}/3

The resultant M by N matrix **ZRI** (*t*) is a smooth estimate of the median home value free of residual systematic error. This may not be as necessary for large regions such as the nation and states because of the large available data set, but it is applied to all levels for consistency.

Final Quality Control

The time series matrix **ZRI**(*t*) has the same dimension as **H**(*t*) which is M by N (as noted, 12 x 57,022). While this theoretically could produce more than 680,000 different time series, in practice many time series are eliminated because of data sparseness or temporal volatility. The general logic determining whether a ZRI time series for a particular combination of region and market segment will be suppressed from the publicly available data set includes the following elements:

1. Number of Rent Zestimates < [threshold]
2. Number of rental listings in most recent three months < [threshold]
3. Temporal volatility measured by annualized, monthly or quarterly change > [threshold]
4. Region has been deemed suspect based on a manual review

Applying the suppression criteria above, there are 195,258 unique deliverable ZRI time series for the report period ending Jan 2012. Table 3 below shows the count of regional time series by region level and market segment. For example, there are 515 time series at the county level for the single-family home variant of the ZRI.

Table 3: Number of deliverable ZRI time series by region level and market segment

Market Segment	National	State	MSA	County	Congressional District	City	Neighborhood	Zip
All Homes	1	43	277	515	352	6,789	4,695	8,916
Single Family	1	43	277	515	352	6,774	4,447	8,838
Condo	1	43	247	424	342	2,970	2,227	4,988
0 or missing	1	43	276	510	352	5,486	2,814	7,595
1 Bedroom	1	43	242	417	346	1,685	970	2,950
3 Bedroom	1	43	275	508	352	5,146	3,055	7,420
2 Bedroom	1	43	277	514	352	6,151	3,921	8,362
4 Bedroom	1	43	276	512	352	5,244	2,723	7,559
5+Bedroom	1	43	268	490	352	2,914	1,377	5,077
Top Tier	1	42	274	497	352	5,510	2,458	7,527
Middle Tier	1	42	274	498	352	6,205	3,339	8,390
Bottom Tier	1	42	274	493	351	5,579	3,512	7,783
Total	**12**	**513**	**3,237**	**5,893**	**4,207**	**60,453**	**35,538**	**85,405**

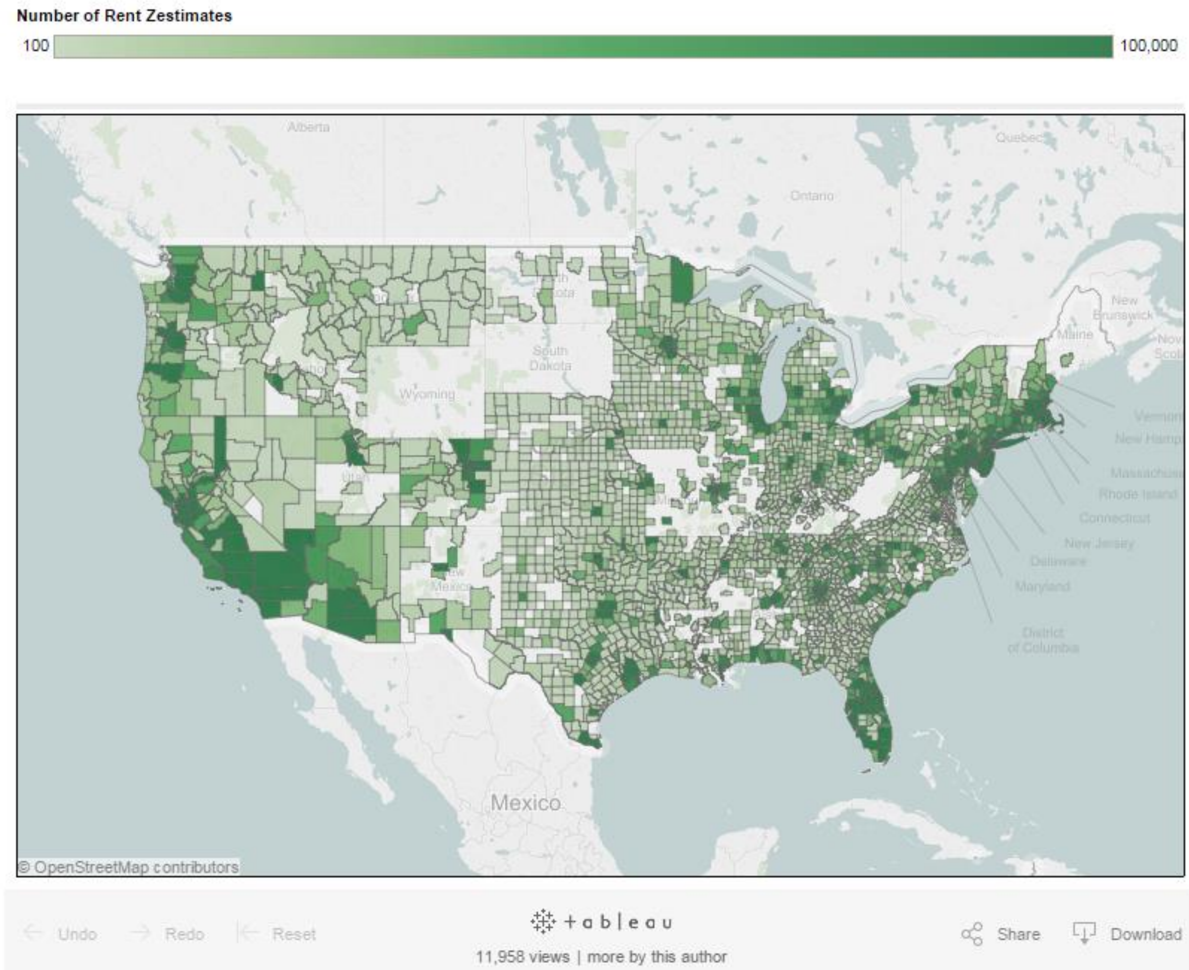
Restatement

Unlike the ZHVI, there is no restatement of the ZRI in the routine monthly calculations because Rent Zestimates do not depend on data that arrive with some latency such as public record transactional data (such as is the case with Zestimates and the corresponding ZHVI). However, there are two situations in which restatements are unavoidable. First, when the boundaries of a geographic region change, the ZRI for the region will change as well since the set of homes underlying the ZRI is different. Second when we regenerate historical Rent Zestimates (for example, when a more accurate algorithm is developed), we also have to re-generate all historical ZRIs.

We are continuously working on improving the underlying algorithm to make Rent Zestimates more accurate. When major improvements to the algorithm are made, we will re-compute the historical Rent Zestimates for affected homes. Our purpose in doing so is to provide consumers with the best estimate of historical rents.

Data Coverage

We calculate the ZRI at the national level as the median Rent Zestimate of 84.9 million homes. The interactive map below displays the number of Rent Zestimates by county for the period ending Jan 31, 2012.

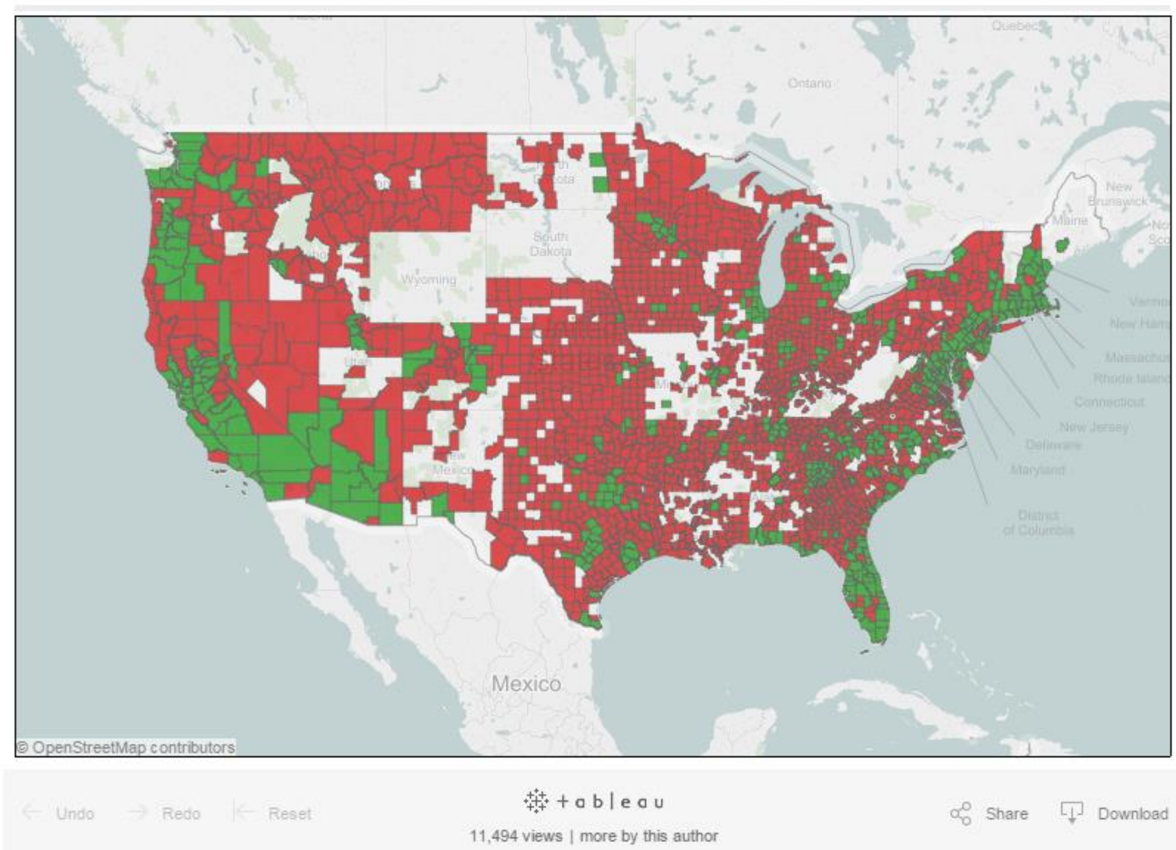


Powered by Tableau

Some county-level ZRIs are suppressed based on the filter rules discussed in the Final Quality Control section above (although Rent Zestimates in those counties are used in computing higher-level ZRIs). The interactive map below shows counties that have a valid ZRI as of January 2012 (green) and those counties where the ZRI has been suppressed based on filter rules but individual Rent Zestimates are still available (red).

ZRI and Rent Zestimates

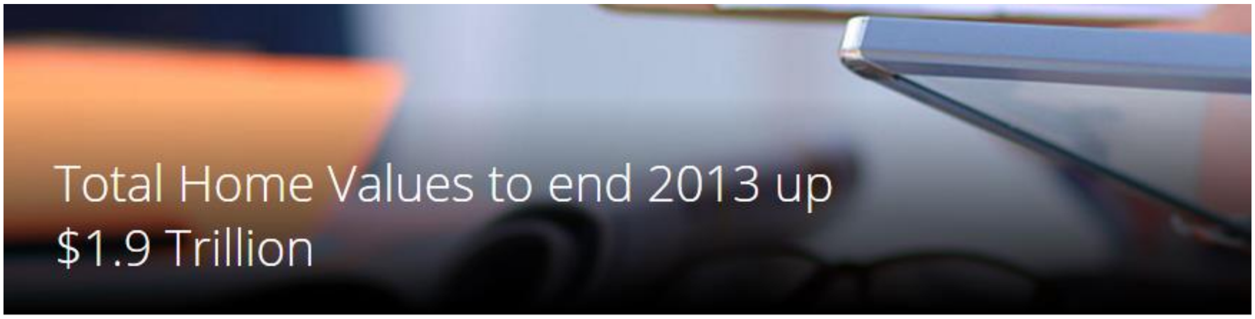
Rent Zestimates only



Powered by Tableau

ABOUT THE AUTHOR

Yeng Bun is a Senior Data Scientist at Zillow.



Total Home Values to end 2013 up $1.9 Trillion

Home / Real Estate Analytics

By Yeng Bun on 12/19/2013

Share

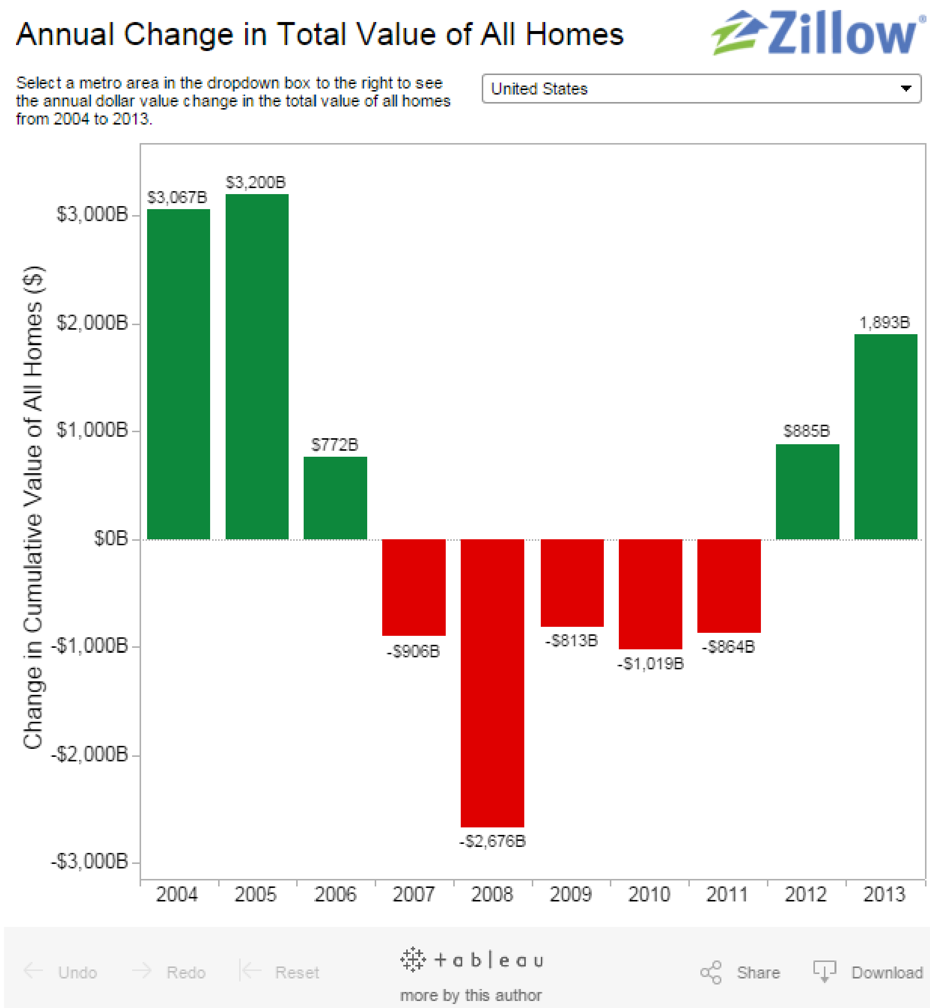
U.S. homes are expected to gain almost $1.9 trillion in cumulative value in 2013, the second annual gain after the housing recession.

Almost 90 percent of the 485 metro areas included in this analysis – 434 in all – experienced cumulative home value gains in 2012. All of the top 30 metro areas covered by this analysis saw gains in the total value of all homes. Of the 30 largest metros, those with the largest gains in cumulative value as measured by total dollar volume include Los Angeles ($323.1 billion), San Francisco ($159.2 billion), New York ($123.1 billion), Miami-Fort Lauderdale ($83.3 billion) and San Diego ($71.5 billion).

Metros that experienced home value losses in 2012 showed strong gains this year. Chicago, Philadelphia and New York all had home value gains in 2013 that offset home value losses in 2012.

Gains were calculated by measuring the difference between cumulative home values as of the end of 2012 and anticipated cumulative home values at the end of 2013. Overall, U.S. homes will have gained approximately $1.89 trillion in cumulative value during full-year 2013, to a total of approximately $25.7 trillion, up 7.9 percent from the end of 2012. Last year, cumulative home values rose for the first time since 2006 and were up $885 billion at the end of 2012.

Home values have been rising for the majority of 2013 before falling in October and November. On an annual basis, home values are up 5.2 percent year over year. Many of the regions that the housing recession hit the hardest have seen especially strong home value appreciation. Among them are California markets, Miami and Phoenix.



ABOUT THE AUTHOR

Yeng Bun is a Senior Data Scientist at Zillow.

Visa Analysis Shows Retail Spending Up in April Across Most Major Categories

05/07/2015
Despite growth in 11 of 14 major purchase categories, a Visa survey finds more than half of consumers say they are pocketing savings from lower gasoline prices

SAN FRANCISCO--(BUSINESS WIRE)-- Visa announced today that Americans modestly increased their spending in April, with growth across most major purchase categories, according to Visa's Retail Spending Monitor (RSM), a quarterly report that tracks retail spending patterns based on real-time purchase data. Retail spending in April was up 4.5 percent from the prior year, excluding automobile and gasoline purchases. Amid a strengthening housing market and renewed confidence in the economy, Americans continue to open their wallets for restaurant meals, hotel stays, household goods like appliances and furniture, and other more day-to-day needs. Eleven of the fourteen major spending categories that Visa tracks showed growth from the prior year.

"Across the country, we're seeing consumers continue to spend as their confidence in the economy grows," said Wayne Best, Visa's Chief Economist. "With spending increases over the prior year from retailers and restaurants as well as a more robust travel sector, this broad-based growth is making an important contribution to the economic recovery."

Several discretionary categories showed solid increases in April from the prior year, with some eclipsing their March growth rates. The increase suggests that American households with incomes greater than $100,000, who generally are more likely to be able to contribute to discretionary spending and less likely to be impacted by swings in gas prices, may be driving the increase in spending. For instance:

Restaurant spending rose 9.5 percent from the prior year in April, compared to a 7.6 percent increase in March.
Hotel spending was up 9.4 percent from the prior year in April, compared to a 9.2 percent increase in March.
Household good spending, at places like electronics, appliance, and furniture stores, increased 5.1 percent in April from the prior year, compared to 1.5 percent in March.

Impact of Gas Prices

Gas prices continue to affect consumers' mind set and spending behavior. Prices have fallen 30 percent over the last year, averaging $2.47 per gallon in April. Consumers received an unexpected windfall on average of $1.19 per gallon compared to a year ago, or between $50 and $75 a month in average household savings.

However, a recent Visa survey found that, amid the increase in gas prices that began in February, more than half of respondents (52 percent) said that they planned to save the unexpected windfall from lower prices at the pump, while nearly a quarter (24 percent) said they planned to use it to pay down debt. Only 30 percent said they planned to spend more at other places.

These survey results are evident in Visa's RSM data. Although April retail spending (excluding autos and gas) was up 4.5 percent from the prior year, it has slowed significantly since the first three months of 2015. In January, when gas prices hit their recent lows, retail spending was up 6.0 percent from the year before. There was also a noticeable impact in consumer spending in several major categories. Some changes include:

- Home improvement spending growth, at places like building supply, hardware, and garden stores, slowed to 4.5 percent in April from the prior year, compared to a 9.4 percent increase in March.
- Clothing store spending increased just 0.1 percent in April from the prior year, after growing by 3.7 percent in March.
- Warehouse and general merchandise spending growth, such as at big-box retailers, slowed to 4.8 percent in April from the prior year, compared to 6.7 percent in March.
- Non-store retail spending growth, such as at online retailers, slowed to 4.6 percent in April from the prior year, compared to 5.5 percent in March.

"What matters is not the price at the pump today, but where consumers see gas prices headed," noted Best. "After gas prices rose every single day in February, 70 percent of consumers said they expected them to keep rising over the next three months – and not surprisingly, they modified their spending habits. We saw that trend again in April when gasoline prices steadily ticked upward in the latter half of the month, causing consumers to spend more cautiously and pocket much of the savings from lower prices at the pump."

About Visa's Retail Spending Monitor

Drawing upon the power of the world's largest payment network, Visa's Retail Spending Monitor provides a real-time window into how and where Americans are spending their money -- and its broader impact on the economy. With billions of transactions flowing through its payment network each day, Visa sees roughly 25 cents of every retail dollar spent in the United States. Using these actual transactions as a starting point, Visa has created a sophisticated, robust model that allows it to gauge overall spending activity across all forms of payment and across major spending categories, including retail, travel and entertainment. The RSM relies on aggregated, depersonalized transaction data and is not reflective of Visa's operational and/or financial performance.

REVIEW SUMMARY

ECONOMICS

Economics in the age of big data

Liran Einav[1,2*] and Jonathan Levin[1,2]

BACKGROUND: Economic science has evolved over several decades toward greater emphasis on empirical work. The data revolution of the past decade is likely to have a further and profound effect on economic research. Increasingly, economists make use of newly available large-scale administrative data or private sector data that often are obtained through collaborations with private firms, giving rise to new opportunities and challenges.

ADVANCES: These new data are affecting economic research along several dimensions. Many fields have shifted from a reliance on relatively small-sample government surveys to administrative data with universal or near-universal population coverage. This shift is transformative, as it allows researchers to rigorously examine variation in wages, health, productivity, education, and other measures across different subpopulations; construct consistent long-run statistical indices; generate new quasi-experimental research designs; and track diverse outcomes from natural and controlled experiments.

Perhaps even more notable is the expansion of private sector data on economic activity. These data, sometimes available from public sources but other times obtained through data-sharing agreements with private firms, can help to create more granular and real-time measurement of aggregate economic statistics. The data also offer researchers a look inside the "black box" of firms and markets by providing meaningful statistics on economic behavior such as search and information gathering, communication, decision-making, and microlevel transactions. Collaborations with data-oriented firms also create new opportunities to conduct and evaluate randomized experiments.

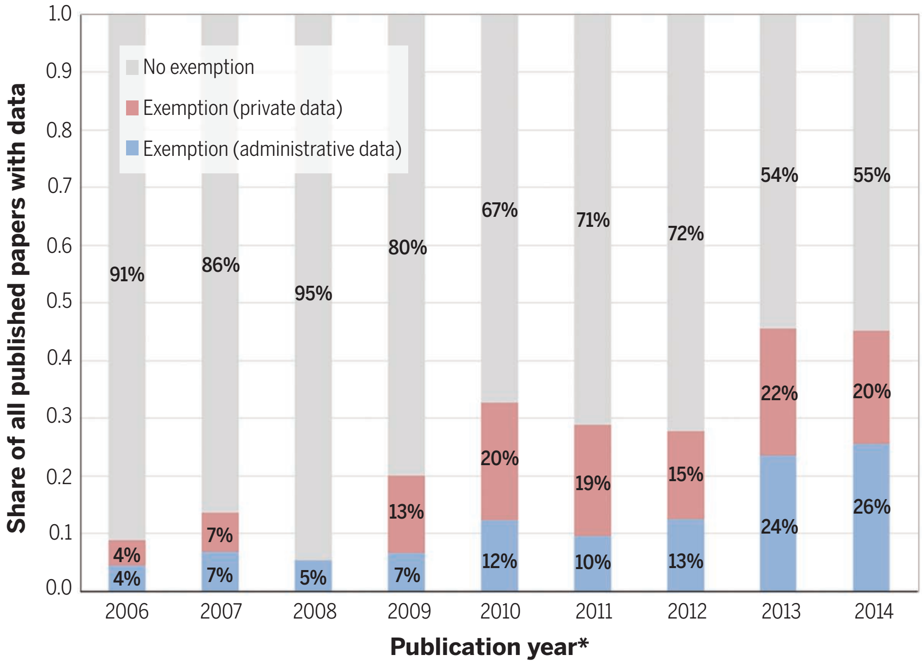
ON OUR WEB SITE

Read the full article at http://dx.doi.org/10.1126/science.1243089

Economic theory plays an important role in the analysis of large data sets with complex structure. It can be difficult to organize and study this type of data (or even to decide which variables to construct) without a simplifying conceptual framework, which is where economic models become useful. Better data also allow for sharper tests of existing models and tests of theories that had previously been difficult to assess.

OUTLOOK: The advent of big data is already allowing for better measurement of economic effects and outcomes and is enabling novel research designs across a range of topics. Over time, these data are likely to affect the types of questions economists pose, by allowing for more focus on population variation and the analysis of a broader range of economic activities and interactions. We also expect economists to increasingly adopt the large-data statistical methods that have been developed in neighboring fields and that often may complement traditional econometric techniques.

These data opportunities also raise some important challenges. Perhaps the primary one is developing methods for researchers to access and explore data in ways that respect privacy and confidentiality concerns. This is a major issue in working with both government administrative data and private sector firms. Other challenges include developing the appropriate data management and programming capabilities, as well as designing creative and scalable approaches to summarize, describe, and analyze large-scale and relatively unstructured data sets. These challenges notwithstanding, the next few decades are likely to be a very exciting time for economic research. ■



The rising use of non–publicly available data in economic research. Here we show the percentage of papers published in the *American Economic Review* (*AER*) that obtained an exemption from the *AER*'s data availability policy, as a share of all papers published by the *AER* that relied on any form of data (excluding simulations and laboratory experiments). Notes and comments, as well as *AER Papers and Proceedings* issues, are not included in the analysis. We obtained a record of exemptions directly from the *AER* administrative staff and coded each exemption manually to reflect public sector versus private data. Our check of nonexempt papers suggests that the *AER* records may possibly understate the percentage of papers that actually obtained exemptions. The asterisk indicates that data run from when the *AER* started collecting these data (December 2005 issue) to the September 2014 issue. To make full use of the data, we define year 2006 to cover October 2005 through September 2006, year 2007 to cover October 2006 through September 2007, and so on.

[1]Department of Economics, Stanford University, Stanford, CA 94305, USA. [2]National Bureau of Economic Research, 1050 Massachusetts Avenue, Cambridge, MA 02138, USA.
*Corresponding author. E-mail: leinav@stanford.edu
Cite this article as L. Einav, J. Levin, *Science* **346**, 1243089 (2014); DOI: 10.1126/science.1243089

REVIEW

ECONOMICS

Economics in the age of big data

Liran Einav[1,2]* and Jonathan Levin[1,2]

The quality and quantity of data on economic activity are expanding rapidly. Empirical research increasingly relies on newly available large-scale administrative data or private sector data that often is obtained through collaboration with private firms. Here we highlight some challenges in accessing and using these new data. We also discuss how new data sets may change the statistical methods used by economists and the types of questions posed in empirical research.

The expansion of data being collected on social and economic activity is likely to have profound effects on economic research. In this Review, we describe how newly available public and private sector data sets are being employed in economics. We also discuss how statistical methods in economics may adapt to take advantage of large-scale granular data, as well as some of the challenges and opportunities for future empirical research.

After providing some brief background in the next section, we divide the Review into three parts. We first discuss the shift from relatively small-sample government surveys to administrative data with universal or near-universal population coverage. These data have been used in Europe for some time but are just starting to be explored in the United States. We explain the transformative power of these data to shed light on variation across subpopulations, construct consistent long-run statistical indices, generate new quasi-experimental research designs, and track diverse outcomes from natural and controlled experiments.

The second part of the Review describes the marked expansion of private sector data on economic activity. We outline the potential of these data in creating aggregate economic statistics and some nascent attempts to do this. We then discuss the rise of collaborations between academics and data-rich companies. These relationships have some trade-offs in terms of maintaining data confidentiality and working with samples that have been collected for business rather than research purposes. But as we illustrate with examples from recent work, they also provide researchers with a look inside the "black box" of firms and markets and create new opportunities to conduct and evaluate randomized experiments.

The third part of this Review addresses statistical methods and the role of economic theory in the analysis of large-scale data sets. Today, economists routinely analyze large data sets with the same econometric methods used 15 or 20 years ago. We contrast these methods to some of the newer data mining approaches that have become popular in statistics and computer science. Economists, who tend to place a high premium on statistical inference and the identification of causal effects, have been skeptical about these methods, which put more emphasis on predictive fit and handling model uncertainty and on identifying low-dimensional structure in high-dimensional data. We argue that there are considerable gains from trade. We also stress the usefulness of economic theory in helping to organize complex and unstructured data.

We conclude by discussing a few challenges in making use of new data opportunities, in particular the need to incorporate data management skills into economics training, and the difficulties of data access and research transparency in the presence of privacy and confidentiality concerns.

The rise of empirical economics

Hamermesh (*1*) recently reviewed publications from 1963 to 2011 in top economics journals. Until the mid-1980s, the majority of papers were theoretical; the remainder relied mainly on "ready-made" data from government statistics or surveys. Since then, the share of empirical papers in top journals has climbed to more than 70%, and a substantial majority of these papers use data that have been assembled or obtained by the authors or generated through a controlled experiment.

This shift mirrors the expansion of available data. Even 15 or 20 years ago, interesting and unstudied data sets were a scarce resource. Gathering data on a specific industry could involve hunting through the library or manually extracting statistics from trade publications. Collaborations with companies were unusual, as were experiments, both in laboratory settings and in the field. Nowadays the situation is very different along all of these dimensions. Apart from simply having more observations and more recorded data in each observation, several features differentiate modern data sets from many used in earlier research.

The first feature is that data are now often available in real time. Government surveys and statistics are released with a lag of months or years. Of course, many research questions are naturally retrospective, and it is more important for data to be detailed and accurate rather than available immediately. However, administrative and private data that are continuously updated have great value for helping to guide economic policy. Below, we discuss some early attempts to use Internet data to make real-time forecasts of inflation, retail sales, and labor market activity and to create new tracking measures of the economy.

The second feature is that data are available on previously unmeasured activities. Much of the data now being recorded is on activities that were previously difficult to quantify: personal communications, social networks, search and information gathering, and geolocation data. These data may open the door to studying issues that economists have long viewed as important but did not have good ways to study empirically, such as the role of social connections and geographic proximity in shaping preferences, the transmission of information, consumer purchasing behavior, productivity, and job search.

Finally, data come with less structure. Economists are used to working with "rectangular" data, with N observations and $K << N$ variables per observation and a relatively simple dependence structure between the observations. New data sets often have higher dimensionality and less-clear structure. For example, Internet browsing histories contain a great deal of information about a person's interests and beliefs and how they evolve over time. But how can one extract this information? The data record a sequence of events that can be organized in an enormous number of ways, which may or may not be clearly linked and from which an almost unlimited number of variables can be created. Figuring out how to organize and reduce the dimensionality of large-scale, unstructured data is becoming a crucial challenge in empirical economic research.

Public sector data: Administrative records

In the course of administering the tax system, social programs, and regulation, the federal government collects highly detailed data on individuals and corporations. The same is true of state and local governments, albeit with less uniformity, in areas such as education, social insurance, and local government spending. As electronic versions of these data become available, they increasingly are the resource of choice for economists who work in fields such as labor economics, public finance, health, and education.

Administrative data offer several advantages over traditional survey data. Workhorse surveys—such as the Survey of Consumer Finances, the Current Population Survey, the Survey of Income and Program Participation, and the Panel Study on Income Dynamics—can suffer from substantial missing data issues, and the sample size may be limited in ways that preclude natural quasi-experimental research designs (*2*). The rich microlevel administrative data sets maintained by, among others, the Social Security Administration, the Internal Revenue Service, and the Centers for Medicare and Medicaid, often have

[1]Department of Economics, Stanford University, Stanford, CA 94305, USA. [2]National Bureau of Economic Research, 1050 Massachusetts Avenue, Cambridge, MA 02138, USA.
*Corresponding author. E-mail: leinav@stanford.edu

high data quality and a long-term panel structure. Sample selection and attrition, a common issue with survey panels, is not a primary concern (3).

These "universal" data sets are especially powerful for analyzing population variation. For instance, Piketty and Saez (4) have used tax records to calculate income and wealth shares for the very upper portion of the income distribution. These calculations are problematic for traditional surveys because of small sample sizes, underreporting of high incomes or asset levels, and the fact that surveys generally extend back only a few years or, at most, decades. In contrast, tax data allow for the creation of relatively homogeneous time series spanning many decades, or even centuries.

Administrative data have been similarly useful in documenting regional disparities in economic mobility (5) (Fig. 1) and health care spending (6), in discovering the wide variation in test-score value-added measures across public school teachers (7), and in identifying the sizable differences in wages and productivity across otherwise similar firms (8, 9). In each case, researchers have used large-scale administrative data to measure and compare the relevant variable (e.g., income, spending, productivity, or wages) across small subpopulations of individuals or firms. These results have helped to guide policy discussions and define research agendas in multiple subfields of economics.

Recent work also highlights the value of using administrative data for causal inference and policy evaluation. For these purposes, administrative data can be valuable both because its coverage and detail allow for novel research designs and because of the possibility of linking records to track outcomes from an existing experiment or quasi-experiment. The last point is an important one. Matching a data set with a random survey of 1 million U.S. households will reduce the original sample to just 1% of its original size. Merging with administrative data may leave the sample virtually unchanged.

Akerman *et al.*'s (10) recent study of the effects of broadband Internet access is illustrative of how administrative data sets can be combined to perform a successful evaluation study. Their research design relies on the gradual expansion of broadband access in Norway into different geographic regions. The authors link this staggered rollout to administrative tax records to estimate how broadband adoption affected firm wages and productivity. By linking individual and firm-level administrative data sets, the authors can observe multiple outcome measures and assess the effect broadband access has on specific subpopulations—for example, broadband access turns out to have very different effects on workers of different education levels.

The same advantages of universal coverage apply when the experiment or quasi-experiment that forms the basis for the study's research design affects only a relatively small population. A recent example is Chetty *et al.*'s (11, 12) study of the long-term effects of teacher quality. The authors use student-level test-score data from a specific city and identify a quasi-experiment in the way students are assigned to teachers that creates variation in teacher quality. The notable step comes when the authors link the student records to administrative tax data and are able to trace the effect of teacher quality on the students' subsequent wages, two decades later.

Several recent studies have also used administrative records in powerful fashion to track outcomes from truly randomized experiments. Chetty *et al.* (13) track the future earnings of students who were randomly assigned to classrooms during the Tennessee STAR (Student-Teacher Achievement Ratio) experiment conducted in the late 1980s. Taubman *et al.*'s (14) evaluation of the Oregon Medicaid expansion similarly uses a range of administrative data to track outcomes after an episode in which Oregon expanded its Medicaid program to a randomly selected subset of newly eligible individuals. The latter study links state administrative data, hospital admission records, private sector credit bureau records, and more targeted survey data to estimate the impact of Medicaid on health and financial measures.

The potential of administrative data for academic research is just starting to be realized, and

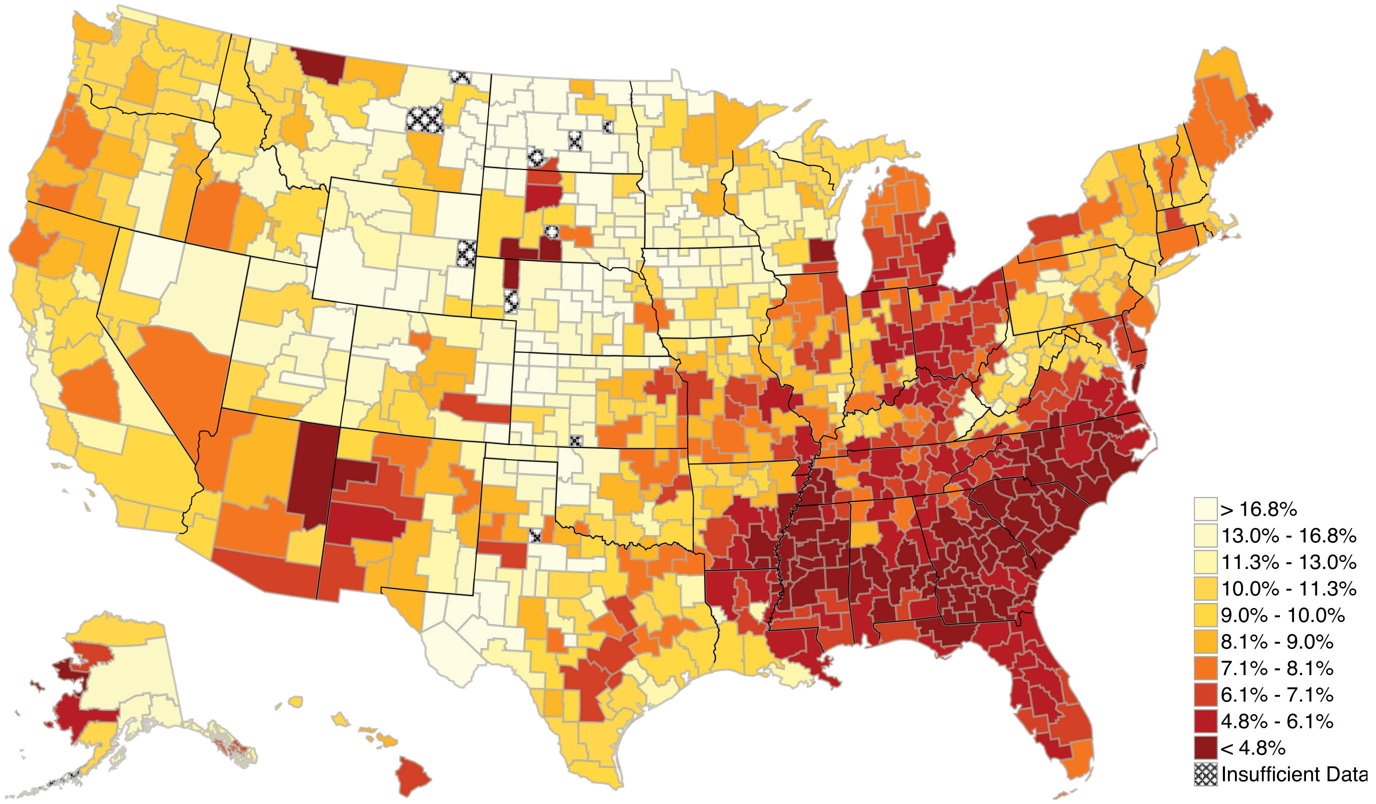


Fig. 1. Economic mobility across U.S. commuting zones. Heat map of upward income mobility using anonymous earnings records on all children in the 1980–1985 birth cohorts. Upward income mobility is measured by the probability that a child reaches the top quintile of the national family income distribution for children, conditional on having parents in the bottom quintile of the family income distribution for parents. Children are assigned to commuting zones based on the location of their parents (when the child was claimed as a dependent), irrespective of where they live as adults. [Reprint of appendix figure VIb in (5)]

substantial challenges remain (15, 16). This is particularly true in the United States, where confidentiality and privacy concerns, as well as bureaucratic hurdles, have made accessing administrative data sets and linking records between these data sets relatively cumbersome. European countries such as Norway, Sweden, and Denmark have gone much farther to merge distinct administrative records and facilitate research. Card *et al.* (3) have articulated a set of principles for expanding access to administrative data, including competition for data access, transparency, and prevention of disclosure of individual records. We view these as useful guideposts. However, even with today's somewhat piecemeal access to administrative records, it seems clear that these data will play a defining role in economic research over the coming years.

Private sector data: Collection and collaborations

An even more dramatic change in data collection is occurring in the private sector. Whereas the popular press has focused on the vast amount of information collected by Internet companies such as Google, Amazon, and Facebook, firms in every sector of the economy now routinely collect and aggregate data on their customers and their internal businesses. Banks, credit card companies, and insurers collect detailed data on household and business financial interactions. Retailers such as Walmart and Target collect data on consumer spending, wholesale prices, and inventories. Private companies that specialize in data aggregation, such as credit bureaus or marketing companies such as Acxiom, are assembling rich individual-level data on virtually every household.

Although the primary purpose of all this data collection is for business use, there are also potential research applications in economics and other fields. These applications are just starting to be identified and explored, but recent research already provides some useful signals of value.

One potential application of private sector data is to create statistics on aggregate economic activity that can be used to track the economy or as inputs to other research. Already the payroll service company ADP publishes monthly employment statistics in advance of the Bureau of Labor Statistics, MasterCard makes available retail sales numbers, and Zillow generates house price indices at the county level. These data may be less definitive than the eventual government statistics, but in principle they can be provided faster and perhaps at a more granular level, making them useful complements to traditional economic statistics.

The Billion Prices Project (BPP) at the Massachusetts Institute of Technology is a related researcher-driven initiative. The BPP researchers coordinate with Internet retailers to download daily prices and detailed product attributes on hundreds of thousands of products (17). These data are used to produce a daily price index. Although the sample of products is, by design, skewed toward products stocked by online retailers, it can replicate quite closely the consumer price index (CPI) series generated by the Bureau of Labor Statistics, with the advantage that the standard consumer series is published monthly, with a lag of several weeks. More interestingly, the project generates price indices for countries in which government statistics are not regularly available or countries in which the published government statistics may be suspect for misreporting, as in Argentina (18) (Fig. 2).

Baker *et al.* (19) have adopted a similar data aggregation strategy by assembling the full texts of 10 leading newspapers to construct a daily index of economic policy uncertainty. In contrast to the BPP indices, their Economic Policy Uncertainty Index is a new measure of economic activity that does not have a parallel in any formal government report. However, it captures a concept that economists have argued may be important for understanding firm investment decisions and macroeconomic activity.

Recent work suggests that publicly available search query data or tweets on Twitter might be used to provide similar statistics on aggregate activity (20, 21). As an example, Varian and coauthors (22, 23) use Google search data to provide short-run forecasts of unemployment, consumer confidence, and retail sales. Their analysis has parallels to the well-known Google Flu Trends index, which used search query data to predict the Center for Disease Control's measure of flu infections. There is a cautionary note here as well, given that the Google Flu Trends index model broke down as Google changed its underlying search algorithm (24). It is likely that successful economic indices using private data will have to be maintained and updated carefully.

A second application of private data is to allow researchers to look "inside" specific firms or markets to study employee or consumer behavior or the operation of different industries. Recent work in this vein often relies on proprietary data obtained through collaborations with private firms. These agreements may take various forms, depending on the sensitivity of the data

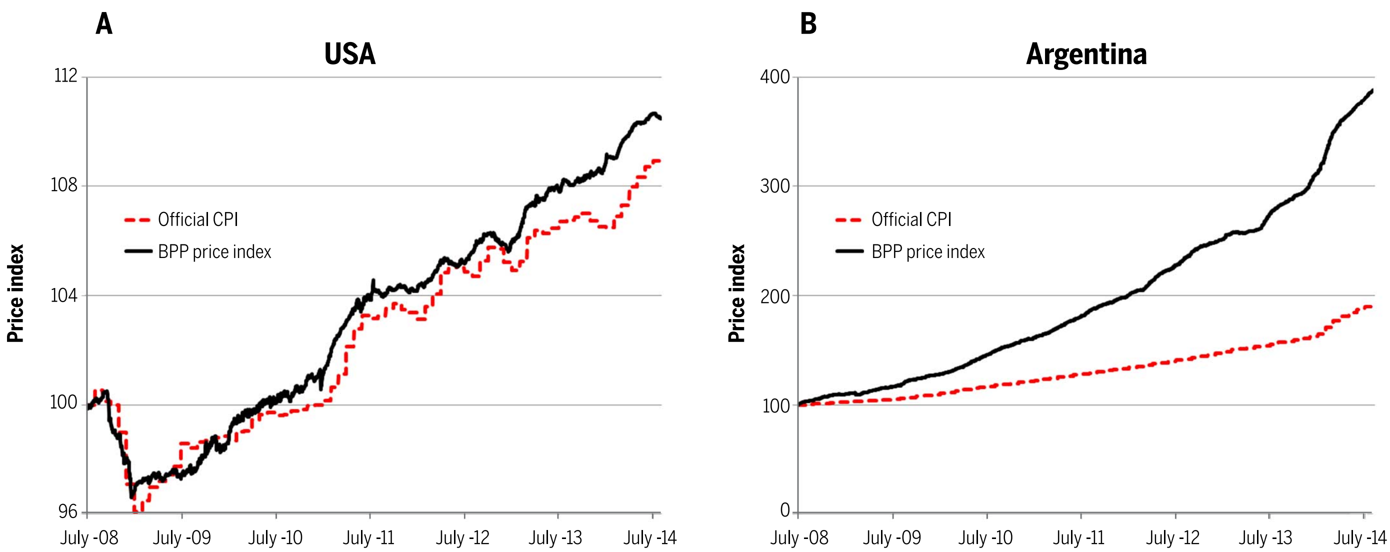


Fig. 2. BPP price index. Dashed red lines show the monthly series for the CPI in the United States (**A**) and Argentina (**B**), as published by the formal government statistics agencies. Solid black lines show the daily price index series, the "State Street's PriceStats Series" produced by the BPP, which uses scraped Internet data on thousands of retail items. All indices are normalized to 100 as of 1 July 2008. In the U.S. context, the two series track each other quite closely, although the BPP index is available in real time and at a more granular level (daily instead of monthly). In the plot for Argentina, the indices diverge considerably, with the BPP index growing at about twice the rate of the official CPI. [Updated version of figure 5 in (18), provided courtesy of Alberto Cavallo and Roberto Rigobon, principal investigators of the BPP]

from a privacy and business perspective. Researchers may have to agree to keep the underlying data confidential. In exchange, however, they often get to work with granular employee- or customer-level data that provide a window into the detailed operations of specific businesses or markets.

Relative to government surveys or administrative data, company data have some important differences. Sampling usually is not representative, and how well findings generalize must be evaluated case by case. Data collection emphasizes recency and relevance for business use, so variables and data collection may not be comparable and uniform over long periods. In short, the data are best viewed as "convenience" samples, albeit with potentially enormous scale. At the same time, private entities are not bound by some of the bureaucratic constraints that limit public agencies. The detail of private data can be much greater, the computing resources can be more powerful, and private companies can have far more flexibility to run experiments.

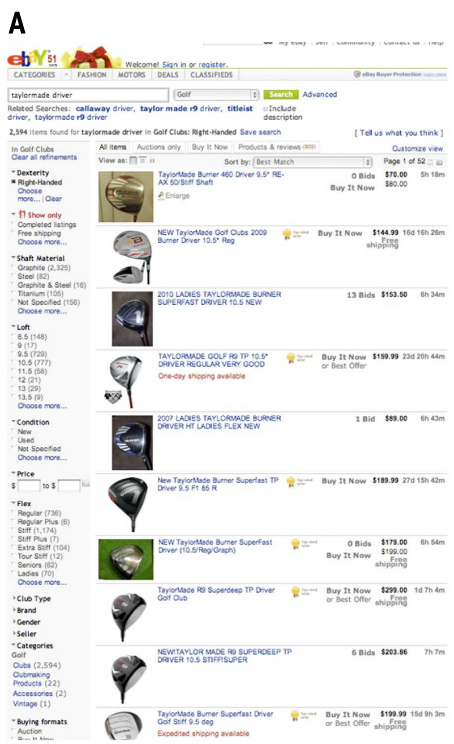
The detail and granularity of private data can offer novel opportunities to study a range of markets. For example, as part of collaboration with researchers at eBay, we recently used their marketplace data to study the effect of sales taxes on Internet shopping (*25*). One of our empirical strategies was to find instances in which multiple consumers clicked on a particular item and then compare consumers located in the same state as the seller (in which case the seller collected sales tax) to consumers located at a similar distance but across state lines (so that no sales tax was collected). The idea of the research design is to assess the sensitivity to sales taxes for otherwise similar consumers looking at the exact same product listing. This sort of analysis would not have been feasible without access to underlying browsing data that allowed us to sift through billions of browsing events to identify the right ones for our empirical strategy.

In two other recent studies (*26*, *27*), also undertaken in collaboration with eBay, we studied the effectiveness of different Internet pricing and sales strategies. To do this, we identified millions of instances in which an online seller listed the same item for sale multiple times with different pricing or shipping fees or using alternative sales mechanisms (e.g., by auction or by posted price) (Fig. 3). We then used the matched listings to estimate the demand response to different item prices and shipping fees, compare auctions with posted price selling, and study alternative sales mechanisms such as auctions with a "buy-now" option. This type of large-scale, microlevel study of market behavior is likely to become more and more common in coming years.

Similar to some of the research described above, a central theme in these papers is the use of highly granular data to find targeted variation that plausibly allows for causal estimates (in these examples, estimates of the effects of sales tax collection, pricing changes, and so forth). In the Internet case, this comes in moving from aggregated data on market prices and quantities to individual browsing data or seller listing data. Having granular data on a market with billions of transactions also provides a chance to analyze specific consumer or market segments: geographic variation, new and used goods, or experienced versus inexperienced sellers. In addition, having richer data can be useful in constructing more nuanced outcome measures. As an example, in studying the effects of sales taxes, we were able to examine not only whether facing a sales tax deterred buyers from purchasing but also whether they continued browsing and then purchased a similar untaxed item.

Large-scale granular data can also be particularly useful for assessing the robustness of identifying assumptions. Virtually every observational study in economics must deal with the critique that even after controlling for sources of confounding, the data do not approximate a controlled experiment. For example, in our work on Internet selling strategies, we aggregated many matched-listing episodes, hoping that each episode might approximate a pricing experiment conducted by the seller. But sometimes sellers may make pricing changes in response to consumer demand, complicating what one can infer from the price change. One way to check if this contaminates the results is to use narrower matching strategies that remove potential sources of confounding—for instance, focusing on cases in which sellers post two offers at the exact same time. This type of extra detective work is much easier with plentiful data.

Collaborations with private sector firms can also give rise to structured economic experiments. This type of research has accelerated



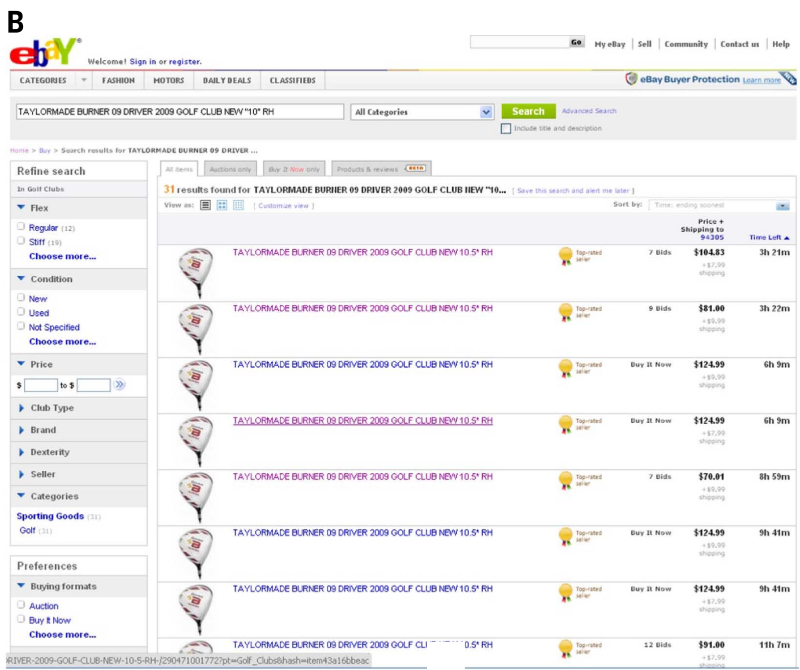


Fig. 3. Matched listings on eBay. (**A**) Screenshot showing a "standard" set of listings on eBay, after a search for "taylormade driver" on 12 September 2010. (**B**) Screenshot showing a matched set. It shows the first 8 out of 31 listings for the same golf driver by the same seller. All of the listings were active on 12 September 2010. Of the eight listings shown, four are offered at a fixed price of $124.99. The other four listings are auctions with slightly varying end times. The listings have different shipping fees (either $7.99 or $9.99). Such matched sets are ubiquitous on eBay and are useful as natural experiments in assessing the effects of changes to sale format and parameters. [Reprint of figure 1 in (*26*)]

and is particularly low-cost and scalable on the Internet, where experimentation is already a standard business practice (28, 29). Recent examples include Ostrovsky and Schwarz (30), who worked with Yahoo! to test the use of different reserve prices in advertising auctions; Blake *et al.* (31), who worked with eBay to selectively shut down its Google search advertising and track the effect on eBay site visits and sales; and Horton (32), who worked with oDesk to provide recommendations to employers about who to hire (33).

As in the case of administrative data, economists working with private companies face some challenges, particularly regarding data access. Although companies may be willing to make small, nonsensitive data sets public, researchers usually have to agree to keep data confidential if they want to work directly with company records. As a result, opportunities for other researchers to replicate or extend studies may be limited. In addition, some collaborative research projects are part of broader consulting or employment relationships, raising issues regarding conflict of interest and selectivity in what results are pursued or submitted for publication.

These issues have only recently become a major topic of discussion in economics, as journals and research organizations have begun to adopt policies on transparency and disclosure. As companies capture increasing amounts of economic data, however, it seems almost certain that collaborations between academics and private sector firms will expand, so we hope that disclosure policies will prove effective and that companies will begin to establish open processes for allowing researchers access to data in ways that reasonably maintain privacy and confidentiality. The underlying issues around data privacy and acceptable types of research experiments are clearly sensitive ones that need to be handled with care and thoughtfulness (34).

Econometrics, machine learning, and economic theory

Recent economic research using large data sets has relied primarily on traditional econometric techniques. The estimated models usually focus on one or a few coefficients of interest, which often represent the causal effect of a particular policy or policies. Researchers put considerable thought and effort into controlling for heterogeneity or other confounding factors, often using a large set of fixed effects, and into obtaining carefully constructed standard errors for the main parameters of interest. Though studies often focus on a single preferred specification, frequently linear, it is typical to assess the robustness of the results by estimating a variety of alternative specifications and running placebo regressions to see if the preferred model generates false-positive findings.

This approach, both in conception and execution, stands in contrast to some of the data mining methods that have become popular for large-data applications in statistics and computer science [e.g., (35, 36)]. These latter approaches put more emphasis on predictive fit, especially out-of-sample fit, and on the use of data-driven model selection to identify the most meaningful predictive variables (37). There often is less attention paid to statistical uncertainty and standard errors and considerably more to model uncertainty. The common techniques in this sort of data mining—classification and regression trees, lasso and methods to estimate sparse models, boosting, model averaging, and cross-validation—have not seen much use in economics (38).

There are some good reasons why empirical methods in economics look the way they do. Economists are often interested in assessing the results of a specific policy or testing theories that predict a particular causal relationship. So empirical research tends to place a high degree of importance on the identification of causal effects and on statistical inference to assess the significance of these effects. Having a model with an overall high degree of predictive fit is often viewed as secondary to finding a specification that cleanly identifies a causal effect.

Consider a concrete example: Suppose we set out to measure whether taking online classes improves a worker's earnings. An economist might hope to design an experiment or to find a natural experiment that induced some workers to take online classes for reasons unrelated to their productivity or current earnings (e.g., a change in the advertising or pricing of online classes). Absent an experimental design, however, she might consider estimating a model such as

$$y_i = \alpha + \beta x_i + \boldsymbol{z}_i'\gamma + \varepsilon_i \qquad (1)$$

where y_i is the outcome (an individual's earnings in a given year), x_i is the policy of interest (whether the worker has taken online classes before that year), β is the key parameter of interest (the effect of online education on earnings), α and γ are other parameters, $\boldsymbol{z}_i$ is a set of control variables, and ε_i is an error term.

The hope is that in a group of individuals with the same $\boldsymbol{z}_i$, whether or not an individual decides to take online classes is not related in a meaningful way to their earnings. Better data obviously help. With detailed individual data over time, the control variables might include a dummy variable for every individual in the sample and perhaps for every employer. Then the effect of online education would be estimated by comparing increases in worker earnings for those who take online classes to increases in earnings for those who do not, perhaps even making the comparison within a given firm. The focus of the analysis would be on the estimate of β, its precision, and on whether there were important omitted variables (e.g., a worker becoming more ambitious and deciding to take classes and work harder at the same time) that might confound a causal interpretation.

Given the same data, a machine learning approach might start with the question of exactly what variables predict earnings, given the vast set of possible predictors in the data, and the potential for building a model that predicts earnings well, both in-sample and out-of-sample. Ultimately, a researcher might estimate a model that provides a way to predict earnings for individuals who have and have not taken online classes, but the exact source of variation identifying this effect—in particular, whether it was appropriate to view the effect as causal—and inference on its statistical significance might be more difficult to assess.

This example may help to illustrate a few reasons economists have not immediately shifted to new statistical approaches, despite changes in data availability. An economist might argue that, short of an experimental approach, the first observational approach has the virtue of being transparent or interpretable in how the parameter of interest is identified, as well as conducive to statistical inference on that parameter. Yet a researcher who wanted to predict earnings accurately might view the first model as rather hopeless, particularly if it included a dummy variable for every individual and the researcher wanted to predict out-of-sample.

However, the two approaches are not necessarily in competition. For instance, if only a subset of control variables is truly predictive, an automated model-selection approach may be helpful to identify the relevant ones (39, 40). Data mining methods may also be useful if there are important interaction effects (41) so that one cares about predicting effects for specific individuals rather than an average effect for the population. A potential benefit of large data sets is that they allow for more tailored predictions and estimates (e.g., a separate β depending on many specifics of the environment). Rather than estimate only average policy treatment effects, it is possible to build models that map individual characteristics into individual treatment effects and allow for an analysis of more tailored or customized policies.

The potential gains from trade go in the other direction as well. To the extent that machine learning approaches are used to assess the effect of specific policy variables and the estimates are given a causal interpretation, the economists' focus on causal identification is likely to be useful.

Economic theory also plays a crucial role in the analysis of large data sets, in large part because the complexity of many new data sets calls for simpler organizing frameworks. Economic models are useful for this purpose.

The connection between big data and economic theory can already be seen in some applied settings. Consider the design of online advertising auctions and exchanges. These markets—run by companies such as Google, Yahoo!, Facebook, and Microsoft—combine big data predictive models with sophisticated economic market mechanisms. The predictive models are used to assess the likelihood that a given user will click on a given ad. This might be enough for a company such as Google or Facebook, with enormous amounts of data, to figure out which ads to show. However, it does not necessarily tell them how much to charge, and given that each ad impression is arguably distinct, trying to experimentally set hundreds of millions of prices could be a challenge. Instead, these companies use (quite sophisticated) auction mechanisms to set prices.

The operation of the auction market depends on the interplay between the predictive modeling and the incentive properties of the auction. Therefore, making decisions about how to run this type of market requires a sophisticated understanding of both big data predictive modeling and economic theory. In this sense, it is no surprise that over the past several years many of the large e-commerce companies have built economics teams (in some cases, headed by high-profile academic researchers) or combined economists with statisticians and computer scientists or that computer science researchers interested in online marketplaces draw increasingly on economic theory.

More generally, we see some of the main contributions that economists can make in data-rich environments as coming from the organizing framework provided by economic theory. In the past century, most of the major advances in economics came in developing conceptual or mathematical models to study individual decisions, market interactions, or the macroeconomy. Frequently, the key step in successful modeling has been simplification: taking a complex environment and reducing it down to relationships between a few key variables. As data sets become richer and more complex and it is difficult to simply look at the data and visually identify patterns, it becomes increasingly valuable to have stripped-down models to organize one's thinking about what variables to create, what the relationships between them might be, and what hypotheses to test and experiments to run. Although the point is not usually emphasized, there is a sense that the richer the data, the more important it becomes to have an organizing theory to make any progress.

Outlook

This review has discussed the ways in which the data revolution is affecting economic and broader social science research. More granular and comprehensive data surely allow improved measurements of economic effects and outcomes, better answers to old questions, and help in posing new questions and enabling novel research designs. We also believe that new data may change the way economists approach empirical research, as well as the statistical tools they employ.

Several challenges confront economists wishing to take advantage of these large new data sets. These include gaining access to data; developing the data management and programming capabilities needed to work with large-scale data sets (*42*); and, most importantly, thinking of creative approaches to summarize, describe, and analyze the information contained in these data (*29*). Big data is not a substitute for common sense, economic theory, or the need for careful research designs. Nonetheless, there is little doubt in our own minds that it will change the landscape of economic research. Here we have outlined some of the vast opportunities. We look forward to seeing how they will be realized.

REFERENCES AND NOTES

1. D. S. Hamermesh, Six decades of top economics publishing: Who and how? *J. Econ. Lit.* **51**, 162–172 (2013). doi: 10.1257/jel.51.1.162
2. R. Chetty, "Time trends in the use of administrative data for empirical research," presentation slides (2012); http://obs.rc.fas.harvard.edu/chetty/admin_data_trends.pdf.
3. D. Card, R. Chetty, M. Feldstein, E. Saez, "Expanding access to administrative data for research in the United States," NSF SBE 2020 white paper ID 112 (2010); www.nsf.gov/sbe/sbe_2020/submission_detail.cfm?upld_id=112.
4. T. Piketty, E. Saez, Inequality in the long run. *Science* **344**, 838–843 (2014). doi: 10.1126/science.1251936; pmid: 24855258
5. R. Chetty, N. Hendren, P. Kline, E. Saez, Where is the land of opportunity? The geography of intergenerational mobility in the United States. *Q. J. Econ.* 10.1093/qje/qju022 (2014). doi: 10.1093/qje/qju022
6. Dartmouth Atlas of Health Care, www.dartmouthatlas.org.
7. S. G. Rivkin, E. A. Hanushek, J. F. Kain, Teachers, schools and academic achievement. *Econometrica* **73**, 417–458 (2005). doi: 10.1111/j.1468-0262.2005.00584.x
8. C. Syverson, What determines productivity? *J. Econ. Lit.* **49**, 326–365 (2011). doi: 10.1257/jel.49.2.326
9. J. M. Abowd, F. Kramarz, D. N. Margolis, High wage workers and high wage firms. *Econometrica* **67**, 251–333 (1999). doi: 10.1111/1468-0262.00020
10. A. Akerman, I. Gaarder, M. Mogstad, "The skill complementarity of broadband Internet," Institute for the Study of Labor (IZA) discussion paper no. 7762 (2013); http://ftp.iza.org/dp7762.pdf.
11. R. Chetty, J. Friedman, J. Rockoff, Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *Am. Econ. Rev.* **104**, 2593–2632 (2014). doi: 10.1257/aer.104.9.2593
12. R. Chetty, J. Friedman, J. Rockoff, Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *Am. Econ. Rev.* **104**, 2633–2679 (2014). doi: 10.1257/aer.104.9.2633
13. R. Chetty *et al*., How does your kindergarten classroom affect your earnings? Evidence from Project Star. *Q. J. Econ.* **126**, 1593–1660 (2011). doi: 10.1093/qje/qjr041; pmid: 22256342
14. S. L. Taubman, H. L. Allen, B. J. Wright, K. Baicker, A. N. Finkelstein, Medicaid increases emergency-department use: Evidence from Oregon's health insurance experiment. *Science* **343**, 263–268 (2014). doi: 10.1126/science.1246183; pmid: 24385603
15. G. King, Ensuring the data-rich future of the social sciences. *Science* **331**, 719–721 (2011). doi: 10.1126/science.1197872; pmid: 21311013
16. H. C. Kum, S. Ahalt, T. M. Carsey, Dealing with data: Governments records. *Science* **332**, 1263 (2011). doi: 10.1126/science.332.6035.1263-a; pmid: 21659589
17. A. Cavallo, "Scraped data and sticky prices," Massachusetts Institute of Technology Sloan working paper no. 4976-12 (2012); www.mit.edu/~afc/papers/Cavallo-Scraped.pdf.
18. A. Cavallo, Online and official price indexes: Measuring Argentina's inflation. *J. Monet. Econ.* **60**, 152–165 (2012). doi: 10.1016/j.jmoneco.2012.10.002
19. S. Baker, N. Bloom, S. Davis, "Measuring economic policy uncertainty," Chicago Booth research paper no. 13-02 (2013); http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2198490.
20. S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, D. J. Watts, Predicting consumer behavior with Web search. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 17486–17490 (2010). doi: 10.1073/pnas.1005962107; pmid: 20876140
21. D. Antenucci, M. Cafarella, M. Levenstein, C. Re, M. Shapiro, "Using social media to measure labor market flows," National Bureau of Economic Research (NBER) working paper no. 20010 (2014); www.nber.org/papers/w20010.
22. H. Choi, H. Varian, Predicting the present with Google trends. *Econ. Rec.* **88**, 2–9 (2012). doi: 10.1111/j.1475-4932.2012.00809.x
23. H. Varian, S. Scott, Predicting the present with Bayesian structural time series. *Int. J. Math. Model. Numer. Optim.* **5**, 4–23 (2014). doi: 10.1504/IJMMNO.2014.059942
24. D. Lazer, R. Kennedy, G. King, A. Vespignani, The parable of Google flu: Traps in big data analysis. *Science* **343**, 1203–1205 (2014). pmid: 24626916
25. L. Einav, D. Knoepfle, J. Levin, N. Sundaresan, Sales taxes and Internet commerce. *Am. Econ. Rev.* **104**, 1–26 (2014). doi: 10.1257/aer.104.1.1
26. L. Einav, T. Kuchler, J. Levin, N. Sundaresan, "Learning from seller experiments in online markets," NBER working paper no. 17385; www.nber.org/papers/w17385.
27. L. Einav, C. Farronato, J. Levin, N. Sundaresan, "Sales mechanisms in online markets: What happened to Internet auctions?" NBER working paper no. 19021 (2013); www.nber.org/papers/w19021.
28. R. Kohavi, R. Longbotham, D. Sommerfield, R. Henne, Controlled experiments on the Web: Survey and practical guide. *Data Min. Knowl. Discov.* **18**, 140–181 (2009). doi: 10.1007/s10618-008-0114-1
29. H. Varian, "Beyond big data," presented at the National Associate for Business Economics Annual Meeting, San Francisco, CA, 7 to 10 September 2013; http://people.ischool.berkeley.edu/~hal/Papers/2013/BeyondBigDataPaperFINAL.pdf.
30. M. Ostrovsky, M. Schwarz, "Reserve prices in Internet advertising auctions: A field experiment," Stanford University Graduate School of Business research paper no. 2054 (2009); http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1573947.
31. T. Blake, C. Nosko, S. Tadelis, "Consumer heterogeneity and paid search effectiveness: A large scale field experiment." NBER working paper no. 20171; www.nber.org/papers/w20171.
32. J. J. Horton, "The effects of subsidizing employer search," New York University working paper (2013); http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2346486.
33. Another example is Lewis and Reiley (*43*), who report on consumer advertising experiments done in conjunction with Yahoo!. These experiments have become common, although Lewis and Rao (*44*) have recently argued that extracting useful information from them may be more challenging than one might have hoped or expected.
34. The recent episode involving an experiment that manipulated Facebook's newsfeed (*45*) is a case in point.
35. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (Springer, New York, 2009).
36. A. Rajaraman, J. D. Ullman, *Mining of Massive Datasets* (Cambridge Univ. Press, New York, 2011).
37. A. Vespignani, Predicting the behavior of techno-social systems. *Science* **325**, 425–428 (2009). doi: 10.1126/science.1171990; pmid: 19628859
38. This statement mainly applies to microeconomics; there is more work in time-series macroeconomics that uses such methods.
39. A. Belloni, D. Chen, V. Chernozhukov, C. Hansen, Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80**, 2369–2429 (2012). doi: 10.3982/ECTA9626
40. A. Belloni, V. Chernozhukov, C. Hansen, "Inference on treatment effects after selection amongst high-dimensional controls," Cemmap working paper no. CWP10/12 (2012); http://arxiv.org/abs/1201.0224.
41. H. Varian, Machine learning: New tricks for econometrics. *J. Econ. Perspect.* **28**, 3–28 (2014). doi: 10.1257/jep.28.2.3
42. D. Lazer *et al*., Computational social science. *Science* **323**, 721–723 (2009). doi: 10.1126/science.1167742 pmid: 19197046
43. R. Lewis, D. Reiley, Online ads and offline sales: Measuring the effects of retail advertising via a controlled experiment on Yahoo! *Quant. Mark. Econ.* **12**, 235–266 (2014). doi: 10.1007/s11129-014-9146-6
44. R. A. Lewis, J. M. Rao, "The unfavorable economics of measuring the returns to advertising," working paper (2014); http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2367103.
45. R. Albergotti, "Facebook experiments had few limits," *Wall Street Journal*, 2 July 2014, http://online.wsj.com/articles/facebook-experiments-had-few-limits-1404344378.
46. L. Einav, J. Levin, "The data revolution and economic analysis," in *Innovation Policy and the Economy*, J. Lerner, S. Stern, Eds. (Univ. of Chicago Press, Chicago, 2014), vol. 14, pp. 1–24.

ACKNOWLEDGMENTS

Parts of this Review draw on an earlier article (*46*). We have benefited from discussions with S. Athey, P. McAfee, and H. Varian. We acknowledge research support from the NSF, the Alfred P. Sloan Foundation, and the Toulouse Network on Information Technology.

10.1126/science.1243089

Economics in the age of big data
Liran Einav and Jonathan Levin
Science **346**, (2014);
DOI: 10.1126/science.1243089

How Individuals Smooth Spending: Evidence from the 2013 Government Shutdown Using Account Data*

Michael Gelman† Shachar Kariv‡ Matthew D. Shapiro§
Dan Silverman ¶ Steven Tadelis‖

February 2015
Revised October 14, 2015

Abstract

Using comprehensive account records, this paper examines how individuals adjusted spending and saving in response to a temporary drop in income due to the 2013 U.S. government shutdown. The shutdown cut paychecks by 40% for affected employees, which was recovered within 2 weeks. Though the shock was short-lived and completely reversed, spending dropped sharply implying a naïve estimate of the marginal propensity to spend of 0.58. This estimate overstates how consumption responded. While many individuals had low liquidity, they used multiple strategies to smooth consumption including delay of recurring payments such as mortgages and credit card balances.

*We thank Kyle Herkenhoff, Patrick Kline, Dimitriy Masterov, Melvin Stephens Jr., and Jeffrey Smith as well as participants at many seminars for helpful comments. This research is supported by a grant from the Alfred P. Sloan Foundation. Shapiro acknowledges additional support from the Michigan node of the NSF-Census Research Network (NSF SES 1131500).

†University of Michigan, mgelman@umich.edu
‡UC Berkeley, kariv@berkeley.edu
§University of Michigan and NBER, shapiro@umich.edu
¶Arizona State University and NBER, Daniel.Silverman.1@asu.edu
‖UC Berkeley, NBER, and CEPR, stadelis@berkeley.edu

"Unless we get our paychecks this coming Monday we don't have the money to cover our mortgage, car payment, and the rest of the bills that we need to pay." — ABC 7 news

1 Introduction

How consumers respond to changes in income is a central concern of economic analysis and is key for policy evaluation. This paper uses the October 2013 U.S. Federal Government shutdown and a newly developed dataset of financial account records to examine how consumers with different levels of liquidity, income, and spending respond to a short-lived and entirely reversed drop in income. For affected government employees, the shutdown caused a sharp decline in income that was recovered within two weeks. The new dataset, derived from the de-identified account records of more than 1 million individuals living in the United States, provides a granular and integrated view of how individuals in different economic circumstances adjusted spending, saving, and debt in response to the shock.

The most important findings are, first, that many workers routinely have very low levels of liquidity, especially in the days just before their regular paycheck arrives. Second, and consistent with low liquidity, spending by affected workers declined sharply in response to the drop in income caused by the shutdown – though the drop lasted at most two weeks and was then offset by an equal increase. Third, the granularity and integration of the data reveal the means used by affected workers to smooth consumption—if not spending—most notably their delay of recurring expenses such as mortgage payments and credit card balances. Last, though many workers found very low-cost ways to weather the shock, some with low liquidity who were already relying on credit card debt accumulated still more credit card debt.

Prior studies that measure the response of individuals to changes in income have faced two challenges. First, the optimal reaction to an income change depends both on whether the change is anticipated, and on its persistence; but standard data sources make it difficult to identify shocks to expected income and the longevity of these shocks. Second, analysis and policy prescriptions often require a comprehensive view of the heterogeneous responses to an income change. Existing data typically capture only some dimensions with sufficient resolution. They may measure total spending with precision, but not savings or debt; or they measure spending and debt well, but do not measure income with similar accuracy.

We overcome the challenge of identifying income shocks and their persistence by using the 2013 U.S. Federal Government shutdown, which produced a significant, temporary, and easily identified negative shock to the incomes of a large number of employees. We address the challenge of measuring a household's full range of responses to this shock by exploiting a new dataset derived from the integrated transactions and balance data of more than 1 million individuals in the U.S.[1]

More specifically, the data allow us to distinguish Federal government employees subject to the shutdown. They are distinguished by the transaction description associated with direct deposit of their paychecks to their bank accounts. Knowing who was subject to the income shock, we can examine their responses in terms of spending and other variables before, during, and after the government shutdown. These responses are estimated by a difference-in-difference approach, where the outcomes of affected government workers are compared with those of a control group consisting of workers that have the same biweekly pay schedule as the Federal government, but who were not subject to the shutdown. The control group is mainly non-Federal workers, though also includes some Federal workers not subject to the shutdown.

The pay of a typical affected worker was 40% below normal during the shutdown because the government was closed from October 1 to October 16, 2013, thus including the last four days of the previous ten-day pay period. By the next pay period, however, government operations had resumed and workers were reimbursed fully for the income lost during the shutdown. The transaction data clearly show this pattern for affected workers. This event combined with the distinctive features of the data, which link income, spending, and liquid assets at a high frequency for each individual, provides an unusual opportunity to study the response to a relatively sizeable shock that affected just the timing of income for individuals across the income distribution, without any net effect on their lifetime incomes. See the related literature section

[1] The data are captured in the course of business by a mobile banking app. While newly developed, this dataset has already proved useful for studying the high-frequency response of spending to regular, anticipated income by levels of spending, income, and liquidity (Gelman et al. 2014). The related literature section below discusses other studies that use similar types of account data.

below for a discussion of the distinctions of this study.

An important fact revealed by the balance records is that many affected employees maintained low levels of liquid assets (checking and saving account balances), especially in the days just before their regular paychecks arrive. Prior to the shutdown, the median worker in the data held an average liquid assets balance sufficient to cover just eight days of average spending. Moreover, liquidity exhibits systematic changes over the pay-cycle. Just before payday, the median level of liquidity is only five days of average spending. Indeed, a substantial fraction of this population barely lives paycheck-to-paycheck. On the day before their paycheck arrives, the bottom third of the liquidity distribution has, on average, a liquid asset balance of zero.

Given such low levels of liquidity, it is perhaps unsurprising that the transaction records show a sharp drop in total spending by affected workers during the week of missing paycheck income. Weekly spending declined by roughly half the reduction in income and then recovered roughly equally over the two pay periods following the end of the shutdown. Econometric analysis reveals a marginal propensity to spend of about 0.58 as a response to the income shock. Most individuals reversed this drop in spending immediately after they receive the paychecks that reimbursed them for their lost income.

On its face, it is troubling that so many affected workers maintained such low liquidity and exhibited such a sharp spending response to an unexpected but brief delay in income. It suggests either that benchmark theories founded on a taste for smoothing consumption are badly specified; that households are inadequately buffered against even very temporary shocks; or that the financial markets that make smoothing possible are functioning poorly.

Further examination of the data reveals, however, that even consumers with low liquidity can smooth consumption better than spending using low-cost methods to shift the timing of payments for committed forms of expenditure. More detailed analysis shows that affected workers delayed mortgage payments, in particular; and many individuals shifted credit card balance payments. At the same time, the data show no increase in spending on credit cards; average debt only increased due to delays in debt payments. Hence, while they responded to the temporary shock by reducing spending, a large part of their reaction was to delay recurring payments that impose little to no penalty. This shows how consumers make use of short-term margins of adjustment that are mostly overlooked in the literature on methods of smoothing consumption in response to at least temporary income shocks. As such, it also reveals a potentially important welfare benefit of, especially, mortgages with low interest rates. Mortgages can function as a (cheap) line of credit that can help smooth even large, if brief, shocks to income at relatively low cost.

While the data show that many affected workers were able to use perhaps unconventional means to smooth consumption, if not spending, for some with low liquidity these methods were either inadequate or unavailable. This group, who was carrying some credit card debt already, emerged from the shutdown with still more debt owing to failure to make payments rather than new borrowing.

The remainder of the paper proceeds as follows. Section 2 describes the paper's relationship to prior studies of individual responses to income shocks. Section 3 provides key facts about the circumstances surrounding the shutdown. Section 4 describes the data and our research design. It establishes that many workers regularly have low liquidity prior to receiving their paycheck. Section 5 estimates the average response of spending and liquid assets to the shock. Section 6 considers heterogeneity in these responses across the liquidity distribution and examines the consequences for credit card debt.

2 Related Literature

The literature concerned with individual responses to income shocks is large. Jappelli and Pistaferri (2010) offer an insightful review. Relative to that large literature, a principal distinction of our paper is the integrated, administrative data that allow accurate observation of liquidity, of the income shock itself, and of several forms of response to the shock. These data thus provide measures of important constraints and outcomes that allow improved inference from the heterogeneous and multi-dimensional reactions to this change in income.

Prior studies of income shocks have mostly relied on the self-reports of survey respondents to provide information either about the shock or about the response of spending and savings and debt.[2] Carroll, Crossley, and Sabelhaus (2013), Dillman and House (2013), Einav and Levin (2014), and others, have called for increased use of administrative records to augment survey research. So far, however, the administrative records available for research have typically represented just a slice of economic activity, either providing information about spending at just one retailer, or about the use of a few credit cards, or about just one form of spending.[3] Other approaches blend survey and administrative data. For example, Broda and Parker (2014) and Parker (2015) use consumer-based scanner data to study the response of spending to an income shock; and use surveys for measuring income.[4] This paper is different from these studies of purely administrative data and from those of blended data sources: The integrated data we use provide an accurate, high-resolution, and high frequency picture of liquidity before the shock, and both the spending and net saving responses to the shock.

By using integrated account records, this paper is part of a new and still small literature that includes Baker (2014), Kuchler (2014), Gelman et al. (2014), and Baker and Yannelis (2015). Baker (2014) uses account records from an online banking app, links them to external data on employers, and instruments for individuals' income changes with news about their employers. Because they are persistent, theory suggests that some of these income shocks (from layoffs or plant closings, e.g.) should have different implications for spending from the one caused by the government shutdown. Nevertheless, like the present paper, Baker (2014) finds evidence of the importance of liquidity (more than debt) for the spending response to an income shock. The present paper is distinct in its study of the methods by which those with very little liquidity smooth consumption through a temporary income shock.

Kuchler (2014) studies integrated account records from an online financial management service that elicits from its customers plans for paying down credit card debt. Kuchler (2014) uses those plans, along with the spending responses to income changes, to evaluate a model of present-biased time preferences.

[2]See, for example, Souleles (1999), Browning and Crossley (2001), Shapiro and Slemrod (2003), and Johnson et al. (2006).

[3]See, for example, Gross and Souleles (2002), Agarwal, Liu, and Souleles (2007), and Aguiar and Hurst (2005).

[4]Einav, Leibtag, and Nevo (2010) discuss the challenges that even scanner technologies like Nielsen Consumer Panel (formerly Homescan), and Feenstra and Shapiro (2003) discuss the challenges of using store-based scanner data to measure expenditure and prices.

Gelman et al. (2014) use a small subsample of the same data we use in this paper to study the spending response to the arrival of predictable (paycheck and Social Security benefit) income. That paper did not examine other outcomes besides spending. Finally, in a complementary study completed shortly after ours, Baker and Yannelis (2015) use data from the same banking app used in Baker (2014) to describe the response of affected government workers to the 2013 shutdown. Baker and Yannelis (2015) focuses on income and spending, but does not integrate those outcomes with financial positions. Their analysis confirms that the spending and income response to the government shutdown is identifiable in these data sets. From these initial facts, their paper analyzes time allocation and home production. Our paper focuses on the precarious liquidity position individuals find themselves in near the end of the paycheck cycle and the different channels through which individuals smoothed their consumption.

The shutdown also is a distinctive shock. The shock is large, negative, and proportional to income. These features stand in contrast to shocks arising from government stimulus payments, which are positive and often weakly related to income. See, for example, Shapiro and Slemrod (1995, 2003), Johnson et al. (2006), Parker et al. (2013), Agarwal et al. (2007), Bertrand and Morse (2009), Broda and Parker (2014), Parker (2015), and Agarwal and Qian (2014). The government shutdown caused a 40% drop in anticipated paycheck income for individuals across a wide range of the income distribution. Thus, unlike the stimulus payments, the shutdown represented a sizeable, albeit temporary, shock even to high-income households.

3 The 2013 U.S. Government Shutdown

3.1 Background

The U.S. government was shut down from October 1 to October 16, 2013 because Congress did not pass legislation to appropriate funds for fiscal year 2014. While Federal government shutdowns have historical precedent, it was difficult to anticipate whether this shutdown would occur and how long it would last.[5] The shutdown was

[5]There have been 12 shutdowns since 1980 with an average length of 4 days. The longest previous shutdown lasted for 21 days in 1995-1996. See Mataconis (2011).

preceded by a series of legislative battles surrounding the Affordable Care Act (ACA), also known as Obamacare. Key events and their timing are described in Figure 1.

Opponents of the ACA in the House of Representatives sought to tie FY 2014 appropriations to defunding the ACA. They used the threat of a shutdown as a lever in their negotiations and thus generated considerable uncertainty about whether a shutdown would occur. Just days before the deadline to appropriate funding and avoid a shutdown, there was substantial uncertainty over what would happen. A YouGov/Huffington Post survey conducted on September 28-29, 2013 showed that 44% of U.S. adults thought Congress would reach a deal to avoid a shutdown while 26% thought they would not, and 30% were unsure. A similar survey taken after the shutdown began on October 2-3, 2013 showed substantial uncertainty over its expected duration. 7% thought the shutdown would last less than a week, 31% thought one or two weeks, 19% thought three or four weeks, and 10% thought the shutdown would last more than a month. 33% were unsure of how long it would last.[6] For most federal employees, therefore, the shutdown and its duration were likely difficult to anticipate at the outset. While it was not a complete surprise, it was a shock to many that the shutdown did indeed occur. On the other hand, as we will discuss in the next subsection, the shutdown was essentially resolved contemporaneously with the receipt of the paycheck affected by the shutdown. Hence, there was no reason based on permanent income to respond to the drop in income.

3.2 Impact on Federal Employees

Our analysis focuses on the consequences of the shutdown for a group of the approximately 2.1 million federal government employees. The funding gap that caused the shutdown meant that most federal employees could not be paid until funding legislation was passed. The 1.3 million employees deemed necessary to protect life and property were required to work. They were not, however, paid during the shutdown for work that they did during the shutdown. The 800,000 "non-essential" employees were simply furloughed without pay.[7] In previous shutdowns, employees were paid

[6]Each survey was based on 1,000 U.S. adults. See YouGov/Huffington Post (2013a, b).

[7]Some federal employees were paid through funds not tied to the legislation in question and were not affected. The Pentagon recalled its approximately 350,000 employees on October 5, reducing the number of furloughed employees to 450,000.

retroactively (whether or not they were furloughed). Of course, it was not entirely clear what would happen in 2013. On October 5, however, the House passed a bill to provide back pay to all federal employees after the resolution of the shutdown. While not definitive, this legislation was strong reassurance that the precedent of retroactive pay would be respected, as in fact it was when the shutdown concluded. After the October 5 Congressional action, most of the remaining income risk to employees was due to the uncertain duration of the shutdown and to potential cost-cutting measures that could be part of a deal on the budget.

Unlike most private sector workers, Federal workers are routinely paid with a lag of about a week, so the October 5 House vote came before reduced paychecks were issued. For most government employees, the relevant pay periods are September 22 - October 5, 2013 and October 6 - October 19, 2013. Because the shutdown started in the latter part of the first relevant pay period, employees did not receive payment for 5 days of the 14 day pay period. For most employees on a Monday to Friday work schedule, this would lead to 4 unpaid days out of 10 working days, so they would receive 40% less than typical pay. The actual fraction varies with hours and days worked and because of taxes and other payments or debits. Since the government shutdown ended before the next pay date, employees who only received a partial paycheck were fully reimbursed in their next paycheck.

Federal government employees are a distinctive subset of the workforce. According to a Congressional Budget Office report (CBO 2012), however, federal employees represent a wide variety of skills and experiences in more than 700 occupations. Compared to private sector employees, they tend to be older, more educated, and more concentrated in professional occupations. Table 1 below reproduces Summary Table 1 in the CBO report. Overall, total compensation is slightly higher for federal employees. Breaking down the compensation difference by educational attainment shows that federal employees are compensated relatively more at low levels of education while the opposite holds for the higher end of the education distribution. In the next section, we make similar comparisons based on Federal versus non-Federal employees in our data. The analysis must be interpreted, however, with the caution that Federal employees may not have identical behavioral responses as the general population. We return to this issue in the discussion of the results.

4 Data and Design

4.1 Data

The source of the data analyzed here is a financial aggregation and bill-paying computer and smartphone application that had approximately 1.5 million active users in the U.S. in 2013.[8] Users can link almost any financial account to the app, including bank accounts, credit card accounts, utility bills, and more. Each day, the application logs into the web portals for these accounts and obtains key elements of the user's financial data including balances, transaction records and descriptions, the price of credit and the fraction of available credit used.

We draw on the entire de-identified population of active users and data derived from their records from late 2012 until October 2014. The data are de-identified and the analysis is performed on normalized and aggregated user-level data as described in the Appendix. The firm does not collect demographic information directly and instead uses a third party business that gathers both public and private sources of demographics, anonymizes them, and matches them back to the de-identified dataset. Appendix Table A (replicated from Table 1 of Gelman et al. 2014) compares the gender, age, education, and geographic distributions in a subset of the sample to the distributions in the U.S. Census American Community Survey (ACS) that is representative of the U.S. population in 2012. The app's user population is not representative of the U.S. population, but it is heterogeneous, including large numbers of users of different ages, education levels, and geographic locations.

We identify paychecks using the transaction description of checking account deposits. Among these paychecks, we identify Federal employees by further details in the transaction description. The appendix describes details for identifying paychecks in general and Federal paychecks in particular. It also discusses the extent to which we are capturing the expected number of Federal employees in the data. The number of federal employees and their distribution across agencies paying them are in line with what one would expect if these employees enroll in the app at roughly the same

[8]We gratefully acknowledge the partnership with the financial services application that makes this work possible. All data are de-identified prior to being made available to the project researchers. Analysis is carried out on data aggregated and normalized at the individual level. Only aggregated results are reported.

frequency as the general population.

4.2 Design: Treatment and Controls

Much of the following analysis uses a difference-in-differences approach to study how Federal employees reacted to the effects of the government shutdown. The treatment group consists of Federal employees whose paycheck income we observe changing as a result of the shutdown. The control group consists of employees that have the same biweekly pay schedule as the Federal government who were not subject to the shutdown (see the Appendix for more details). The control group is mainly non-Federal employees, but also includes some Federal employees not subject to the shutdown.[9] Table 2 shows summary statistics from the app's data for these groups of employees. As in the CBO study cited above, Federal employees in our sample have higher incomes. They also have higher spending, higher liquid balances, and higher credit card balances.

We use the control group of employees not subject to the shutdown to account for a number of factors that might affect income and spending during the shutdown: these include aggregate shocks and seasonality in income and spending. Additionally, interactions of pay date, spending, and day of week are important (see Gelman et al. 2014). Requiring the treatment and control to have the same pay dates and pay date schedule (biweekly on the Federal schedule) is a straightforward and important way to control for these substantial, but subtle effects.

There is, of course, substantial variability in economic circumstance across individuals both within and across treatment and controls. We normalize many variables by average daily spending, or where relevant by average account balances) at the individual level. This normalization is a simple, and given the limited covariates in the data, a practical way to pool individuals with very different levels of income and

[9]Employees not subject to the shutdown include military, some civilian Defense Department, Post Office, and other employees paid by funds not involved in the shutdown. An alternative strategy would use just the Federal workers not affected by the shutdown as the control. We do not adopt this strategy because we believe a priori that it is less suitable because the workers exempted from the shutdown are in non-random agencies and occupations. This selectivity makes them potentially less suitable as a control. For completeness, however, the Appendix includes key results using the unaffected Federal workers as the control. The findings are quite similar though less precise because of smaller sample size.

spending. In particular, it serves to equalize the differences in income levels between treatment and control seen in Table 2.

By showing a wide span of data before and after the shutdown, Figure 2 provides strong evidence of the adequacy of the control group and the effectiveness of using average daily spending as the normalization. Figure 2 shows that the employees not subject to the shutdown have nearly identical movement in spending except during the weeks surrounding the shutdown. Thus the controls appear effective at capturing aggregate shocks, seasonality, payday interactions, etc. In particular, note the regular, biweekly pattern of fluctuations in spending. It arises largely from the timing of spending following receipt of the bi-weekly paychecks. There are also subtler beginning-of-month effects—also related to timing of spending. In subsequent figures we use a narrower window to highlight the effects of the shutdown.

Gelman et al. (2014) shows that much of the sensitivity of spending to receipt of paycheck, like that seen in Figure 2, arises from reasonable choices of individuals to time recurring payments—such as mortgage payments, rent, or other recurring bills—immediately after receipt of paychecks. Figure 2 makes clear that the control group does a good job of capturing this feature of the data and therefore eliminating ordinary paycheck effects from the analysis. The first vertical line in Figure 2 indicates the week in which employees affected by the shutdown were paid roughly 40% less than their average paycheck. There is a large gap between the treatment and control group during this week. Similarly, the second vertical line indicates the week of the first paycheck after the shutdown. The rebound in spending is discernable for two weeks. The figure thus demonstrates that the control group represents a valid counterfactual for spending that occurred in the absence of the government shutdown.

4.3 Liquidity Before the Shutdown

To understand how affected employees responded to the shutdown, it is useful to examine first how they and others like them managed their liquid assets prior to the shock. Analysis of liquid asset balances before the shutdown shows that, while some workers were well buffered, many were ill-prepared to use liquid assets to smooth even a brief income shock.

We define liquid assets as the balance on all checking and savings accounts. The measure of liquidity is based on daily snapshots of account balances. Hence, they are measures of the stock of liquid assets independent from the transactional data used to measure spending and income. Having such high-frequency data makes it possible to observe distinctive, new evidence on liquidity and how it interacts with shocks. Figure 3 shows median liquidity over the pay-cycle, by terciles of the distribution of liquidity. Liquidity is expressed as a ratio of checking and savings account balances to average daily total spending. The results are for the period prior to the shutdown and aggregate over both treatment and control groups.[10,11]

While the optimal level of liquidity is not clear, the figure shows the top third of the liquidity distribution is well-positioned to handle the income shock due to the shutdown. The median of this group could maintain more than a month of average spending with their checking and savings account balances, even in the days just before their paycheck arrives.

The lower two-thirds of the liquidity distribution has a substantially smaller cushion. Over the entire pay-cycle, the middle tercile has median liquid assets equal to 7.9 days of average spending. Liquidity drops to only 5 days of average spending in the days just before their paycheck arrives. Thus, even in the middle of the liquidity distribution many would be hard pressed to use liquid savings to smooth a temporary loss of 4 days pay. The bottom third of this population is especially ill-prepared. Prior to the shutdown, the median of this group consistently arrives at payday with precisely zero liquid balances. (Balances can be negative owing to overdrafts.) These balance data thus reveal how, even among those with steady employment, large fractions of consumers do not have the liquid assets to absorb a large, but brief, shock to income.

[10]The distributions for treatment and control are similar. For example, in the control group the median liquidity ratio for the first, second, and third, terciles of the liquidity distribution is, 2.9, 7.9, and 32.1, respectively. The analogous numbers for the treatment group are 3.3, 8.1, and 32.0.

[11]Liquidity peaks two days after a payday. The balance data are based on funds available, so liquidity should lag payday according to the banks funds-availability policy. There is at least one-day lag built into the data because the balances are scraped during the day, so will reflect a paycheck posted the previous day. Appendix Figure A4 shows that the two-day delay in the peak of liquidity is due to funds availability, not delays in posting based on interactions of day of payday and delays in posting of transactions over the week-end. (As discussed in the Appendix, even within the government bi-weekly pay schedule, there is some heterogeneity in day of week of the payday.) Additionally, liquidity is, of course, net of inflows and outflows. Recurring payments made just after the receipt of paycheck will therefore lead daily balances to understate gross liquidity right after the receipt of the paycheck.

5 Responses to the Shutdown

Having established that many (affected) workers had little liquidity prior to the shutdown, we now examine how their income, and various form of spending responded to the shutdown. Our method is to estimate the difference-in-difference, between treatment and control, for various outcomes using the equation,

$$y_{i,t} = \sum_{k=1}^{T} \delta_k \times Week_{i,k} + \sum_{k=1}^{T} \beta_k \times Week_{i,k} \times Shut_i + \Gamma' X_{i,t} + \epsilon_{i,t}, \tag{1}$$

where y represents the outcome variable (total spending, non-recurring spending, income, debt, savings, etc.), i indexes individuals ($i \in \{1, ..., N\}$), and t indexes time ($t \in \{1, ..., T\}$). $Week_{i,k}$ is a complete set of indicator variables for each individual-week in the sample, $Shut_i$ is a binary variable equal to 1 if individual i is in the treatment group and 0 otherwise, and $X_{i,t}$ represents controls to absorb the predictable variation arising from bi-weekly pay week patterns.[12] The β_k coefficients capture the average weekly difference in the outcome variables of the treatment group relative to the control group. Standard errors in all regression analyses are clustered at the individual level and adjusted for conditional heteroskedasticity.

5.1 Paycheck Income and Total Spending

We begin with an examination of how income, as measured in these data, was affected by the shutdown. External reports indicate that the paycheck income of affected employees should have dropped by 40% on average. The analysis of paycheck income here can thus be viewed, in part, as testing the ability of these data to accurately measure that drop. Once that ability is confirmed, we move to an evaluation of the spending responses.

Recall that we normalize each variable of interest, measured at the individual level, by the individual's average daily spending computed over the entire sample period. The unit of analysis in our figures is therefore days of average spending. Figure 4 plots the estimated β_k from equation (1) where y is normalized paycheck income.

[12] Specifically, $X_{i,t}$ contains dummies for paycheck week, treatment, and their interaction. This specification allows the response of treatment and control to ordinary paychecks to differ. These controls are only necessary in the estimates for paycheck income.

We plot three months before and after the government shutdown to highlight the effect of the event. The first vertical line (dashed-blue) represents the week that the shutdown began and the second vertical line (solid-red) represents the week in which pay dropped due to the shutdown, and the third when pay was restored.

Panel A of Figure 4 shows, as expected, a drop in income equal to approximately 4 days of average daily total spending during the first paycheck period after the shutdown.[13] This drop quickly recovers during the first paycheck period after the shutdown ends, as all users are reimbursed for their lost income. Some users received their reimbursement paychecks earlier than usual, so the recovery is spread across two weeks. The results confirm that the treatment group is indeed subject to the temporary loss and subsequent recovery of income that was caused by the government shutdown, and that the account data allow an accurate measure of those income changes.

Panel B of Figure 4 plots the results on total spending, showing the estimated β_k where y is normalized total spending. On average, total spending drops by about 2 days of spending in the week the reduced paycheck was received. Hence, the drop in spending upon impact is about half the drop in income. That implies a propensity to spend of about one-half—much higher than most theories would predict for a drop of income that was widely expected to be made up in the relatively near future. In the inter-paycheck week, spending is about normal. In the second week after the paycheck affected by the shutdown, spending rebounds with the recovery spread mainly over that week and the next one.

To ease interpretation we convert the patterns observed in Figure 4 into an estimate of the marginal propensity to spend (which we call the MPC as conventional). Let τ be the week of the reduced paycheck during the shutdown. The variable $s_{i,\tau-k}$ denotes total spending for individual i in the k weeks surrounding that week. To estimate the MPC, we consider the relationship,

$$s_{i,\tau-k} = \alpha_k + \beta_k^{MPC}(Paycheck_{i,\tau} - Paycheck_{i,\tau-2}) + \epsilon_{i,\tau-k}, \tag{2}$$

[13] The biweekly paychecks dropped by 40 percent on average. For the sample, paycheck income is roughly 70 percent of total spending on average because there are other sources of income. So a drop of paycheck income corresponding to 4 days of average daily spending is about what one would expect (4 days $\approx 0.4 \times 0.7 \times 14$ days).

where $(Paycheck_{i,\tau} - Paycheck_{i,\tau-2})$ is the change in paycheck income. Both $s_{i,\tau-k}$ and $Paycheck_{i,\tau}$ are normalized by individual-level average daily spending as discussed above. We present estimates for the one and two week anticipation of the drop in pay ($k = 1$ and $k = 2$), the contemporaneous MPC ($k = 0$), and one lagged MPC ($k = -1$). We do not consider further lags because the effect of the lost pay is confounded by the effect of the reimbursed pay beginning at time $\tau + 2$.

There are multiple approaches to estimating equation (2). The explanatory variable is the change in paycheck. We are interested in isolating the effect on spending due to the exogenous drop in pay for employees affected by the shutdown. While in concept this treatment represents a 40 percent drop in income for the affected employees and 0 for the controls, there are idiosyncratic movements in income unrelated to the shutdown. First, not all employees affected by the shutdown had exactly a 40 percent drop in pay because of differences in work schedule or overtime during the pay period. Second, there are idiosyncratic movements in pay in the control group. Therefore, to estimate the effect of the shutdown using equation (2) we use an instrumental variables approach where the instrument is a dummy variable $Shut_i$. The IV estimate is numerically equivalent to the difference-in-difference estimator.[14]

Table 3 shows the estimates of the MPC. These estimates confirm that the total spending of government employees reacted strongly to their drop in income and that this reaction was focused largely during the week that their reduced paycheck arrived. The estimate of the average MPC is 0.58 in this week, with much smaller coefficients in the two weeks just prior. Thus, at the margin, about half of the lost income was reflected in reduced spending.

[14] Estimating equation (2) by least squares should produce a substantially attenuated estimate relative to the true effect of the shutdown if there is idiosyncratic movement in income among the control group, some of which results in changes in spending. In addition, if the behavioral response to the shutdown differs across individuals in ways related to variation in the change in paycheck caused by the shutdown (e.g., because employees with overtime pay might have systematically different MPCs), the difference between the OLS and IV estimates would also reflect treatment heterogeneity. This heterogeneity could lead the OLS estimate to be either larger or smaller than the IV estimate, depending on the correlation between of the size of the shutdown-induced shock and the MPC. The OLS estimate of the MPC for the week the reduced paycheck arrived is 0.123, with a standard error of 0.004.

5.2 Spending and Payments by Type

Analyzing different categories of spending offers further insight into the response of these users to the income drop. We separate spending into non-recurring and recurring components. Recurring spending is identified using patterns in both the amount and transaction description of each individual transaction.[15] It identifies spending that, due to its regularity, is very likely to be a committed form of expenditure (see Grossman and Laroque (1990), Chetty and Szeidl (2007), and Postlewaite, et al. (2008)). Non-recurring spending is total spending minus recurring spending. These measures thus use the amount and timing of spending rather than an a priori categorization based on goods and services. This approach to categorization is made possible by the distinctive features of the data infrastructure.

Figure 5 presents estimates of the β_k from equation (1) where the outcome variable y takes on different spending, payment, or transfer categories. For each graph, the data are normalized by individual-level averages for the series being plotted. In the top two panels we can compare the normalized response of recurring and non-recurring spending and see important heterogeneity in the spending response by category. The results on total spending (Figure 4) showed an asymmetry in the spending response before and after the income shock; total spending dropped roughly by 2 days of average spending during the three weeks after the shutdown began and only rose by 1.6 days of average spending during the three weeks after the shutdown ended. The reaction of recurring spending drives much of that asymmetry; it dropped by 2.6 days of average recurring spending and rose only by 0.84 days once the lost income was recovered. Non-recurring spending exhibits the opposite tendency: it dropped by 1.8 days of average non-recurring spending and rose by 2.0 days. Thus, recurring spending drops more and does not recover as strongly as non-recurring spending.

To better understand this pattern of recurring expenditure and its significance we focus on a particular, and especially important, type of recurring spending—mortgage payments. Panel C of Figure 5 shows that, while the mortgage spending data is noisier

[15] We identify recurring spending using two techniques. First, we define a payment as recurring if it takes the same amount at a regular periodicity. This definition captures payments such as rent or mortgages. Second, we also use transaction fields to identify payments that are made to the same payee at regular intervals, but not necessarily in the same amount. This definition captures payments such as phone or utility bills that are recurring, but in different amounts. See appendix for further details. Gelman et al. (2014) uses only the first technique to define recurring payments.

than the other categories, there is a significant drop during the shutdown and this decline fully recovers in the weeks when the employees' missing income was repaid. In this way, we see that some users manage the shock by putting off mortgage payments until the shutdown ends. Indeed, many of those affected by the shutdown changed from paying their mortgage early in October to later in the month as shown in Figure 6. The irregular pattern of payment week of mortgage reflects interaction of the bi-weekly paycheck schedule with the calendar month. The key finding of this figure is that the deficit in payments of the treatment group in the second week of October is largely offset by the surplus of payments in the last two weeks of October.

Panel D of Figure 5 shows the response of account transfers to the income shock.[16] During the paycheck week affected by the shutdown, transfers fell and rebounded when the pay was reimbursed two weeks later. This finding implies a margin of adjustment, reducing transfers out of linked accounts, during the affected week. One might have expected the opposite, i.e., an inflow of liquidity from unlinked asset accounts to make up for the shortfall in pay. That kind of buffering is not present on average in these data.

Similar behavior is seen in the management of credit card accounts. Another relatively low-cost way to manage cash holdings is to postpone credit card balance payments. Panel E of Figure 5 shows there was a sharp drop in credit card balance payments during the shutdown, which was reversed once the shutdown ended. For users who pay their bill early, this is an easy and cost-free way to finance their current spending. Even if users are using revolving debt, the cost of putting off payments may be small if they pay off the balance right away after the shutdown ends. We examine credit card balances in greater detail in the next section.

Indeed, as we see in Panel F of Figure 5, there was no average reaction of credit card spending to the shutdown. Thus, we find no evidence that affected employees sought to fund more of their expenditure with credit cards but instead floated, temporarily, more of their prior expenditure by postponing payments on credit card balances. Affected individuals who had ample capacity to borrow in order to smooth

[16] These are transactions explicitly labeled as "transfer," etc. For linked accounts, they should net out (though it is possible that a transfer into and out of linked accounts could show up in different weeks). Hence, these transfers are (largely) to and from accounts (such as money market funds) that are not linked.

spending, by charging extra amounts to credit cards, had other means of smoothing, e.g., liquid checking account balances or the postponement of mortgage payments. On the other hand, those who one might think would use credit cards for smoothing spending because they had little cash on hand did not—either because they were constrained by credit limits or preferred to avoid additional borrowing. In the next section we will examine the consequences for credit card balances of these postponed balance payments, and later probe the heterogeneous responses of individuals by their level of liquidity.

This analysis of different categories of spending reveals that users affected by the shutdown reduce spending more heavily on recurring spending and payments compared to non-recurring spending. It is important to note that this behavior appears to represent, in many cases, a temporal shifting of payments and neither a drop in eventual spending over a longer horizon or a proportionate drop in contemporaneous consumption. These results thus provide evidence of the instruments that individuals use to smooth temporary shocks to income that has not been documented before. The drop in non-recurring spending shows, however, that this method of cash management is not perfect; it does not entirely smooth spending categories that better reflect consumption.

Spending could have fallen in part because employees stayed home and engaged in home production instead of frequenting establishments that they encounter during their work-day. Recall, however, that many employees affected by the shutdown were not, in fact, furloughed. They worked but did not get paid for that work on the regular schedule. In addition, Figure 7 shows that categories of expenditure that are quite close to consumption, such as a fast food and coffee shops spending index, show a sharp drop during the week starting October 10 when employees were out of work. Given that a cup of coffee or fast food meal is non-durable, one would not expect these categories to rebound after the shutdown. Yet, there is significant rebound after the shutdown. We interpret this spending as resulting from going for coffee, etc., with co-workers after the shutdown, perhaps to trade war stories.[17] Hence, in a sense, a cup of coffee is not entering the utility function as an additively separate non-durable.

[17] Interestingly, the rebound is highest for the most liquid individuals (figure not reported) who are also higher income. This finding supports our notion that the rebound in coffee shop and fast food arises from post-shutdown socializing.

5.3 Response of Liquid Assets

For users who have built up a liquid asset buffer, they may draw down on these reserves to help smooth income shocks. Figure 8 shows the estimated β_k from equation (1) where $y_{i,t}$ is the weekly average liquid balance normalized by its individual-level average (Panel A) or normalized by individual-level average daily total spending (Panel B). Because of the heterogeneity in balances, normalizing by average liquid balances leads to more precise results. Normalizing by total spending is less precise but allows for a more meaningful interpretation because it is in the same units as Figure 4. Consistent with the spending analysis, relative savings for the treatment group rises in anticipation of the temporary drop in paycheck income. There is a steep drop in the average balance the week of the lower paycheck as a result of the shutdown. The drop in liquidity is, however, substantially attenuated relative to the drop in income because of the drop in payments that is documented in the previous section. The recovery of the lost income causes a large spike in the balances, which is mostly run off during the following weeks. Figure 8B shows that liquid balances fell by around 2 day of average daily total spending. Therefore, on average, users reduced spending by about 2 days and drew down about 2 days of liquidity to fund their consumption when faced with a roughly 4 days drop in income. These need not add up because of transfers from non-linked accounts and because of changes in credit card payments, though they do add up roughly at the aggregate level. In the next section, we explore the heterogeneity in responses as a function of liquid asset positions where specific groups of individuals do use other margins of adjustment than liquid assets.

6 Liquidity and the Heterogeneity in Response to the Income Shock

The preceding results capture average effects of the shutdown. There are important reasons to think, however, that different employees will react differently to this income shock, depending on their financial circumstances. Although all may have a desire to smooth their spending in response to a temporary shock, some may not have the means to do so.

In this section we examine the heterogeneity in the response along the critical dimension of liquidity. For those with substantial liquid balances relative to typical spending, it should be relatively easy to smooth through the shutdown. Section 4 showed, however, that many workers in these data have little liquidity, especially in the days just before their regular paycheck arrives. For those (barely) living paycheck-to-paycheck, even this brief drop in income may pose significant difficulties.

We investigate the impact of the shutdown for those with varying levels of liquidity by first further quantifying the buffer of liquid assets that different groups of workers had. Second, we return to each of the spending categories examined above and compare how different segments of the liquidity distribution responded to the income shock. Last, we study how the precise timing of the shock, relative to credit card due dates, influenced credit card balances coming out of the shutdown.

6.1 Liquidity and Spending

As before, we define the liquidity ratio for each user as the average daily balance of checking and savings accounts to the user's average daily spending until the government shutdown started on October 1, 2013 and then divide users into three terciles. Table 4 shows characteristics of each tercile. Users in the highest tercile have on average 54 days of daily spending on hand while the lowest tercile only has about 3 days. This indicates that a drop in income equivalent to 4 days of spending should have significantly greater effects for the lowest tercile compared with the highest tercile.

Figure 9 plots the estimates of β_ks from equation (1), for various forms of spending, by terciles of liquidity. The results are consistent with liquidity playing a major role in the lack of smoothing. Users with little buffer of liquid savings are more likely to have problems making large and recurring payments such as rent, mortgage, and credit card balances. In terms of average daily expenditure, spending for these recurring payments drops the most for low liquidity users. In contrast, the drop in non-recurring spending is similar across all liquidity groups. Like those with more liquidity, however, low liquidity users refrained from using additional credit card spending to smooth the income drop.

6.2 Liquidity and Credit Debt

The preceding results indicate that the sharp declines in recurring spending (especially mortgages) and credit card balance payments induced by the shutdown were particularly important strategies for those with lower levels of liquidity. The granularity of the data shows, however, that fine differences in timing are consequential when liquidity levels are so low.

To examine how individuals manage credit card payments and balances, we carry out the analysis at the level of the individual credit card account, rather than aggregating across accounts as in the previous section. The account-level analysis allows us to examine the role of payment due dates in the response to the shutdown. These due dates may represent significant requirements for liquidity. That they are staggered and unlikely to be systematically related to the timing of the shutdown provides another means for identifying behavioral responses that exploits the high resolution of the data infrastructure.

In this analysis, however, attention is restricted to the accounts of "revolvers." We focus, that is, on accounts held by those who, at some point during the study period (including the period of the shutdown), incurred interest charges on at least one of their credit cards, indicating that they carried some revolving credit card debt. This represents 63% of the treatment group and 63% of controls; and 70% of these workers fall in the lower two-thirds of the liquidity distribution. The complement of the revolver group is the "transactors." Members of this group routinely pay their entire credit card balance, and have a distinct monthly pattern of balances that reflects their credit card spending over the billing cycle and regular payment of the balance at the end of the cycle. Only 44% of transactors fall in the lower two thirds of the liquidity distribution. Including transactors would obscure the results for those who carry credit card debt.[18]

Figure 10 shows the response of credit card balances, at the account level, to the loss of income due to the shutdown. The estimates again present the difference-in-difference between accounts held by revolvers in the treatment group and those held

[18] We investigated those who shifted from being transactors to revolvers at the time of the shutdown. This group was so small (17%) that it did not yield interesting results. Given that transactors tend to have high liquidity (15.9 median ratio vs 7.7 for revolvers), the lack of such transitions is not surprising.

by revolvers in the control group. These estimates are specified in terms of days since the account's August 2013 statement date instead of calendar time in order to show the effect of statement due dates. In Figure 10, Days 0 through 30 on the horizontal axis correspond to payment due dates in late August or in September 2013. (Payments are due typically 25 days after the statement date.) The different panels of Figure 10 show alternative cuts of the data that we will explain next. Focus, however, on the first 25 days since the August statement date, i.e., due dates that occur in advance of the shutdown. Regardless of cut on the data, the difference-in-difference between treatment and control is essentially zero.

Panels A and B divide the sample of accounts into two groups based on the credit card statement date and, in particular, whether the statement date places them "at risk" for having to make a payment during the government shutdown. Panel A shows the accounts with statement dates on September 16-30, 2013. Panel B shows accounts that have statement dates on September 1-15. For those in the treatment group, the accounts with September 16-30 due dates (Panel A) are at risk. Based on our analysis of liquidity over the paycheck cycle (Figure 3) it is likely that the mid-October paycheck that is diminished by the shutdown would have been a primary source of liquidity for making the payment on these accounts that come due during that pay period. Indeed, Panel A reveals this effect. Control and treatment accounts start to diverge about a week to 10 days into the October billing cycle (days 35-37). By the time the November statement arrives (days 58-60), a significant gap emerges; relative to controls, treatment account balances are now significantly above average. They return to average in a month, presumably as affected workers use retroactive pay to make balance payments. Panel B, those who made their payments before the shutdown, shows no such effect (the hump starting at day 30 is prior to the shutdown and is not statistically significant.)

The high-resolution analysis made possible with the data infrastructure reveals that, when liquidity is so low, small differences in timing can matter. Workers whose usual credit card payment date fell before the shutdown adjusted on other margins; their balances did not rise. For others, the shutdown hit just as they would have normally made their credit card payment; they deferred credit card payments and their balances were elevated for a billing cycle or two before returning to normal levels.

These findings for credit cards reinforce the findings for mortgage payments found in the previous section and Figure 6. For those who typically made payments on mortgages early in the month, that is, prior to the receiving the paycheck reduced by the shutdown, there is little effect of the shutdown on mortgage payments. For those who make payments in the second half of the month, they can and often did postpone the mortgage payments as a way to respond to the shock to liquidity.

7 Conclusion

Living paycheck-to-paycheck lets workers consume at higher levels, but would seem to leave them quite vulnerable to income shocks. The results of this paper reveal how workers use financial assets and markets, sometimes in unconventional ways, to reduce that vulnerability and adjust to shocks when they do occur. The findings indicate that to the extent a large but brief shock to income is a primary risk, a lack of liquid assets as a buffer is not necessarily a sign of myopia or unfounded optimism. Rather, the reactions to the 2013 government shutdown studied in this paper indicate that workers can defer debt payments and thus maintain consumption (at low cost) despite limited liquid assets. They may face higher costs to access less liquid assets. Such illiquidity may be optimal even if it leads to short- or medium-run liquidity constraints (see Kaplan and Violante 2013). This paper shows that the majority of households have such liquidity constraints, yet they have mechanisms for coping with shocks to income so as to mitigate the consequences of such illiquidity.

This paper provides direct evidence on the importance of deferring debt payments, especially mortgages, as an instrument for consumption smoothing. Mortgages function for many as a primary line of credit. By deferring a mortgage payment, they can continue to consume housing, while waiting for an income loss to be recovered. For changing the timing of mortgage payments within the month due, there is no cost. As discussed above, that is the pattern for the bulk of deferred mortgage payments. Moreover, the cost of paying one month late can also be low. Many mortgages allow a grace period after the official due date, in which not even late charges are incurred, or charge a fee that is 4-6 percent of the late payment. Being late by a month adds only modestly to the total mortgage when interest rates are low, and many mortgage service companies will not report a late payment to credit agencies until it is at least 30 days overdue. Even if there are penalties or costs, late payment of mortgage is a source of credit that is available without the burden of applying for credit.

Thus, this paper's findings indicate that policies that encourage homeownership and low-interest mortgages may have under-appreciated welfare benefits to those mortgage holders. Our results suggest expansion of mortgage availability not only finances housing, but has the added effect of making it easier to smooth through shocks to income. As in Herkenhoff and Ohanian (2013), who show how skipping mortgage payments can function as a form of unemployment insurance, the results here reveal how the ability to defer mortgage payments can be an important source of consumption insurance in the face of large, temporary income shocks.

The timing of credit card balance payments provides another source of managing liquidity to buffer consumption against a temporary decline in income. For those with low levels of liquid assets, deferring or reducing credit card payments is a convenient and relatively low-cost way to address a temporary income shortfall. Among credit card borrowers who had payment due dates during the pay period with the reduced paycheck, we see significant deferral of payments. Their credit card balances rose, and stayed elevated for a billing cycle or two before returning to normal.

The distinctive findings of this paper derive high-frequency data on transactions and balances that provide new and distinctive evidence on consumer behavior. The precision and resolution of these data allow insights into behavior that are obscured by conventional data sources.

References

Agarwal, Sumit, Chunlin Liu, and Nicholas S. Souleles (2007) "The Reaction of Consumer Spending and Debt to Tax Rebates: Evidence from Consumer Credit Data," *Journal of Political Economy*, **115(6)**: 986-1019.

Agarwal, Sumit, and Wenlan Qian (2014) "Consumption and Debt Response to Unanticipated Income Shocks: Evidence from a Natural Experiment in Singapore," *American Economic Review*, **104(12)**: 4205-30.

Aguiar, Mark and Erik Hurst (2005) "Consumption versus Expenditure," *Journal of Political Economy*, **113(5)**: 919-948.

Baker, Scott (2014) "Debt and the Consumption Response to Household Income Shocks," manuscript, Kellogg School of Management.

Baker, Scott, and Constantine Yannelis (2015) "Income Changes and Consumption: Evidence from the 2013 Federal Government Shutdown." SSRN Working paper, March 2015.

Bertrand, Marianne, and Adair Morse (2009) "Indebted Households and Tax Rebates," *American Economic Review*, **99(2)**: 418-23.

Broda, Christian and Jonathan A. Parker (2014) "The Economic Stimulus Payments of 2008 and the Aggregate Demand for Consumption," *Journal of Monetary Economics*, **68(S)**: S20-S3.

Carroll, Christopher, Thomas Crossley, and John Sabelhaus (2013) *Improving the Measurement of Consumer Expenditures.* University of Chicago Press.

Chetty, Raj and Adam Szeidl (2007) "Consumption Commitments and Risk Preferences," *Quarterly Journal of Economics*, **122(2)**: 831-877.

Congressional Budget Office (2012) "Comparing the Compensation of Federal and Private-Sector Employees."

Dillman, Don A. and Carol C. House eds. (2013) *Measuring What We Spend: Toward a New Consumer Expenditure Survey.* The National Academies Press.

Einav, Liran, Ephraim Leibtag, and Aviv Nevo (2010) "Recording discrepancies in Nielsen Homescan data: Are they present and do they matter?" *Quantitative Marketing and Economics*, **8(2)**: 207-239.

Einav, Liran and Jonathan D. Levin (2014) "Economics in The Age of Big Data," *Science*, **346(6210)**.

Feenstra, Robert C. and Matthew D. Shapiro eds. (2003) *Scanner Data and Price Indexes.* Chicago: University of Chicago Press.

Gelman, Michael, Shachar Kariv, Matthew D. Shapiro, Dan Silverman, and Steven Tadelis (2014) "Harnessing Naturally Occurring Data to Measure the Response of Spending to Income," *Science*, **345(6193)**: 212-215.

Gross, David B. and Nicholas S. Souleles (2002) "Do Liquidity Constraints and Interest Rates Matter for Consumer Behavior? Evidence from Credit Card Data," *The Quarterly Journal of Economics*, **117(1)**: 149-185.

Grossman, Sanford. J. and Guy Laroque (1990) "Asset Pricing and Optimal Portfolio Choice in the Presence of Illiquid Durable Consumption Goods," *Econometrica*, **58(1)**: 25-51.

Herkenhoff, Kyle F., and Lee E. Ohanian (2013) "Foreclosure Delay and U.S. Unemployment," Working paper, University of Minnesota Department of Economics.

Jappelli, Tullio and Luigi Pistaferri (2010) "The Consumption Response to Income Changes," *Annual Review of Economics*, **2**: 479-506.

Johnson, David S., Jonathan A. Parker and Nicholas S. Souleles (2006) "Household Expenditure and the Income Tax Rebates of 2001," *American Economic Review*, **96(5)**: 1589-1610.

Kaplan, Greg, and Gianluca Violante. (2014) "A Model of the Consumption Response to Fiscal Stimulus Payments" *Econometrica*, **82(4)**: 1199-1239

Kuchler, Theresa (2014) "Sticking to Your Plan: Hyperbolic Discounting and Credit Card Debt Paydown," manuscript, NYU Stern School of Business.

Mataconis, Doug (2011) "A Brief History Of Federal Government Shutdowns" Outside the Beltway, 08 April 2011. Web. 10 Feb 2015. http://www.outsidethebeltway.com/a-brief-history-of-federal-government-shutdowns/.

Parker, Jonathan (2015) "Why Don't Households Smooth Consumption? Evidence from a 25 Million Dollar Experiment," unpublished paper, MIT.

Parker, Jonathan A., Nicholas S. Souleles, David S. Johnson, and Robert McClelland (2013) "Consumer Spending and the Economic Stimulus Payments of 2008," *American Economic Review*, **103(6)**: 2530-53.

Postlewaite, Andrew, Larry Samuelson, and Dan Silverman (2008) "Consumption Commitments and Employment Contracts," *Review of Economic Studies*, **75(2)**: 559-578.

Shapiro, Matthew D., and Joel Slemrod (1995) "Consumer Response to the Timing of Income: Evidence from a Change in Tax Withholding," *American Economic Review*, **85(1)**: 274-283.

Shapiro, Matthew D., and Joel Slemrod (2003) "Consumer Response to Tax Rebates," *American Economic Review,* **93(1)**: 381-396

Souleles, Nicholas S. (1999) "The Response of Household Consumption to Income Tax Refunds," *American Economic Review*, **89(4)**: 947-958.

YouGov/Huffington Post (2013a) "Do you think President Obama and Republicans in Congress will reach a deal to avoid a government shutdown, or that they will not reach a deal and the government will shut down?" Question 4 of 6, September 28 – September 29, 2013. Retrieved from https://today.yougov.com/news/2013/09/30/poll-results-shutdown/.

YouGov/Huffington Post (2013b) "How long do you think the government shutdown will last before President Obama and Congress reach a new budget deal?" Question 5 of 5, October 2 – October 3, 2013. Retrieved from https://today.yougov.com/news/2013/10/04/poll-results-federal-shutdown-hp/.

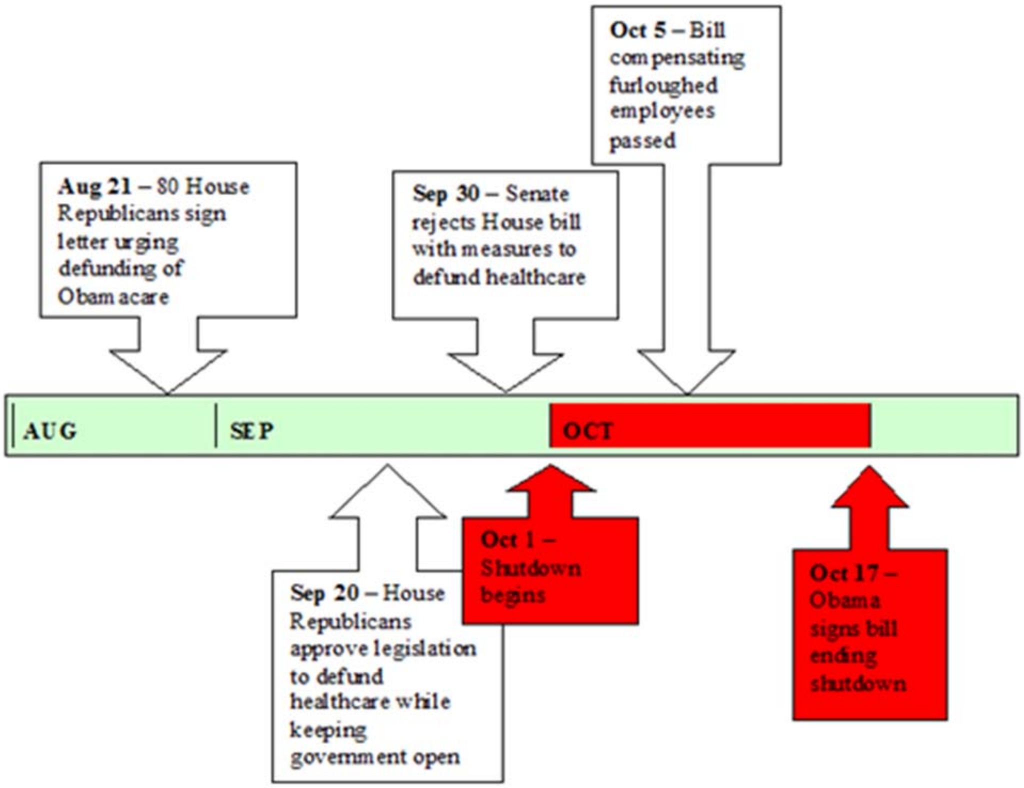


FIGURE 1. GOVERNMENT SHUTDOWN TIMELINE

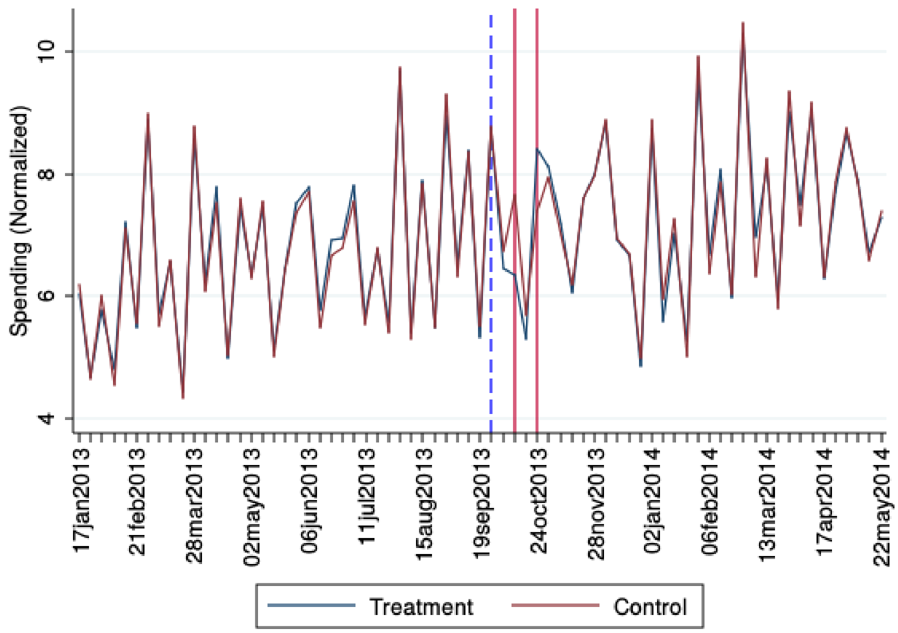


FIGURE 2. TIME SERIES OF SPENDING

Notes: The figure shows average weekly spending (normalized by individuals' daily spending over the entire sample) for government employees subject to the shutdown (treatment) and other employees on the same biweekly pay schedule (control). The first vertical line is the week in which paychecks were reduced owing to the shutdown. The second vertical line indicates the week where government most affected employees received retroactive pay.

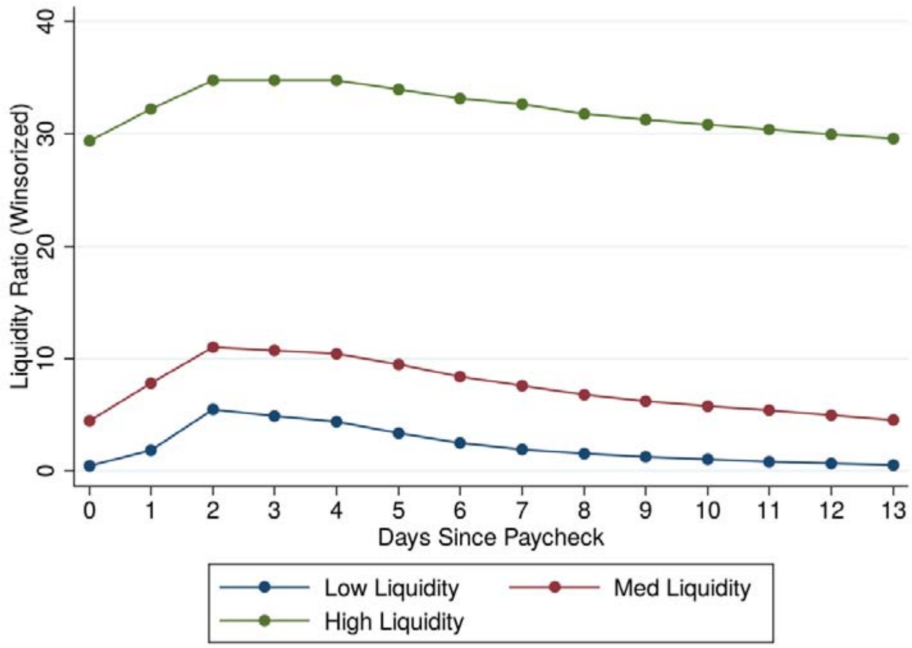


FIGURE 3. PRE-SHUTDOWN MEDIAN LIQUIDITY OVER THE PAYCHECK CYCLE

Notes: Liquidity ratio is defined as checking and savings account balances normalized by average daily spending. The figure shows median liquidity ratio with terciles by days since receipt of paycheck.

A. Paycheck Income B. Total Spending

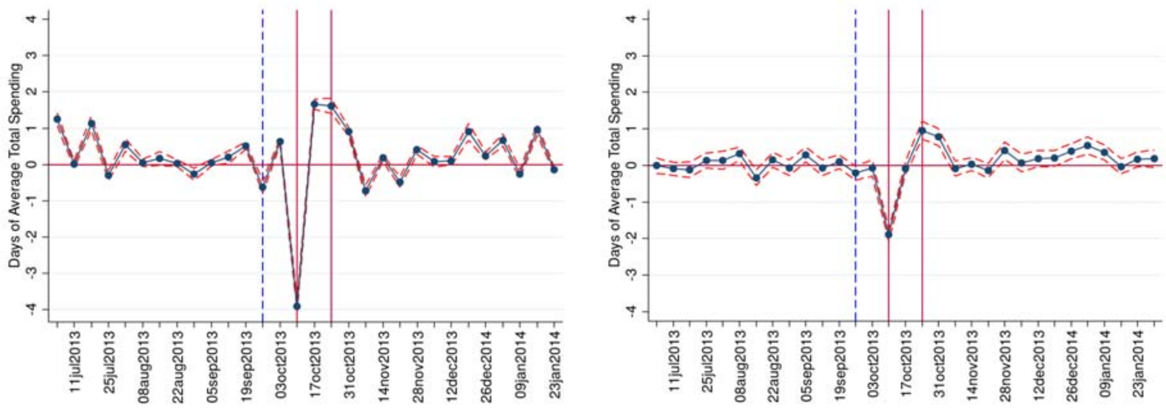


FIGURE 4. ESTIMATED RESPONSE OF NORMALIZED PAYCHECK INCOME AND NORMALIZED TOTAL SPENDING TO GOVERNMENT SHUTDOWN

Notes: Difference-in-difference estimates based on equation (1). Both paycheck income and total spending are normalized by individual-level average daily total spending. The paycheck income plot is estimated using additional controls which include paycheck week and treatment group interactions. N = 3,804 and N= 94,680 for treatment and control group respectively. The estimation period is January 17, 2013 to May 22, 2014. The figures, however, display only the period from July 4, 2013 to January 30, 2014.

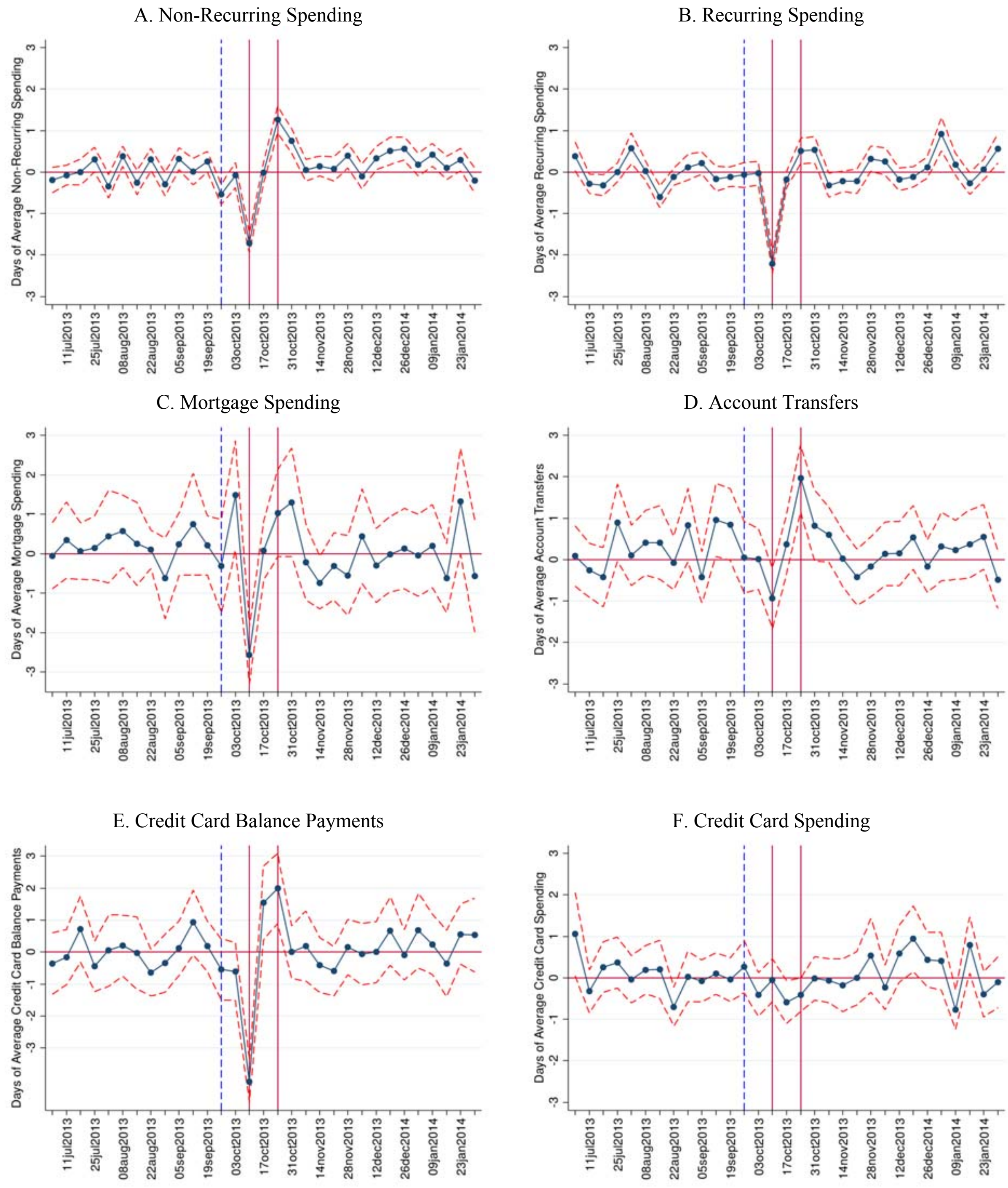


FIGURE 5. ESTIMATED RESPONSE OF SPENDING CATEGORIES TO GOVERNMENT SHUTDOWN

Notes: The spending, payment, or transfer category in each panel is normalized by the individual-level daily average for that category. N = 3,804 and N= 94,680 for treatment and control group respectively.

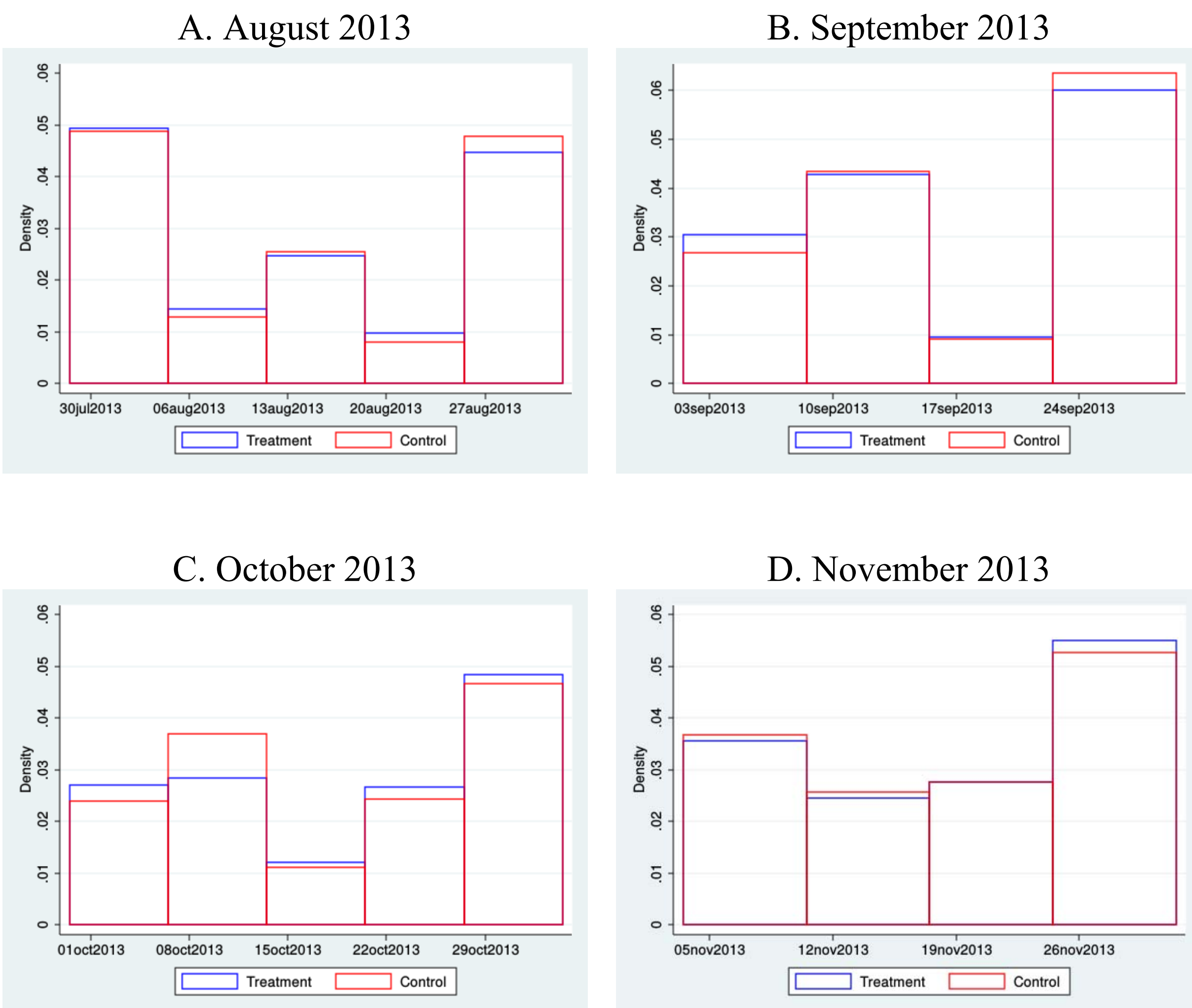


FIGURE 6. DISTRIBUTION OF WEEK MORTGAGE IS PAID

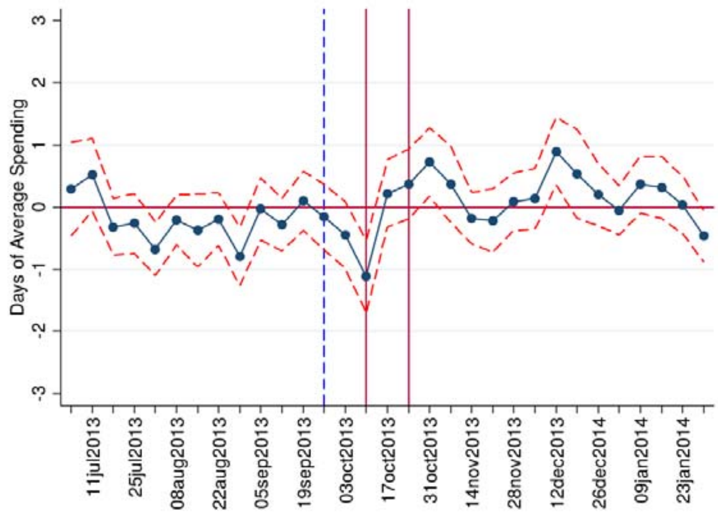


FIGURE 7. ESTIMATED RESPONSE OF COFFEE SHOP AND FAST FOOD SPENDING TO GOVERNMENT SHUTDOWN

Notes: Normalized by individual-level average daily coffee shop and fast food spending. N = 3,804 and N= 94,680 for treatment and control group respectively.

A. Normalized by Average Liquid Balance B. Normalized by Average Daily Spending

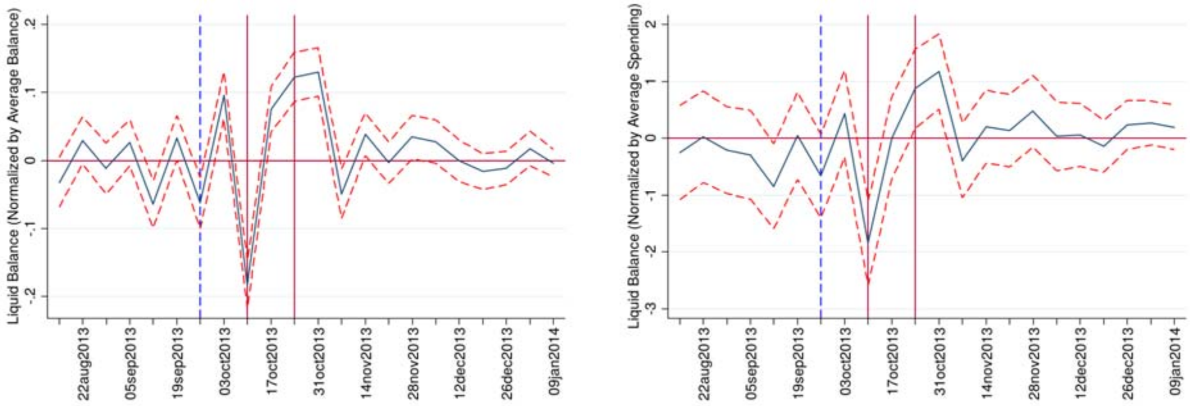


FIGURE 8. ESTIMATED RESPONSE OF LIQUID ASSETS TO GOVERNMENT SHUTDOWN

Notes: Panel A shows end-of-week liquidity (checking plus saving balances) normalized by individual-level average liquidity. Panel B shows end-of-week liquidity normalized by individual-level average daily total spending (same normalization as Figure 4). The treatment group includes 3,804 individuals and the control group includes 94,669 individuals. Outcome variables are winsorized at the upper and lower 1%.

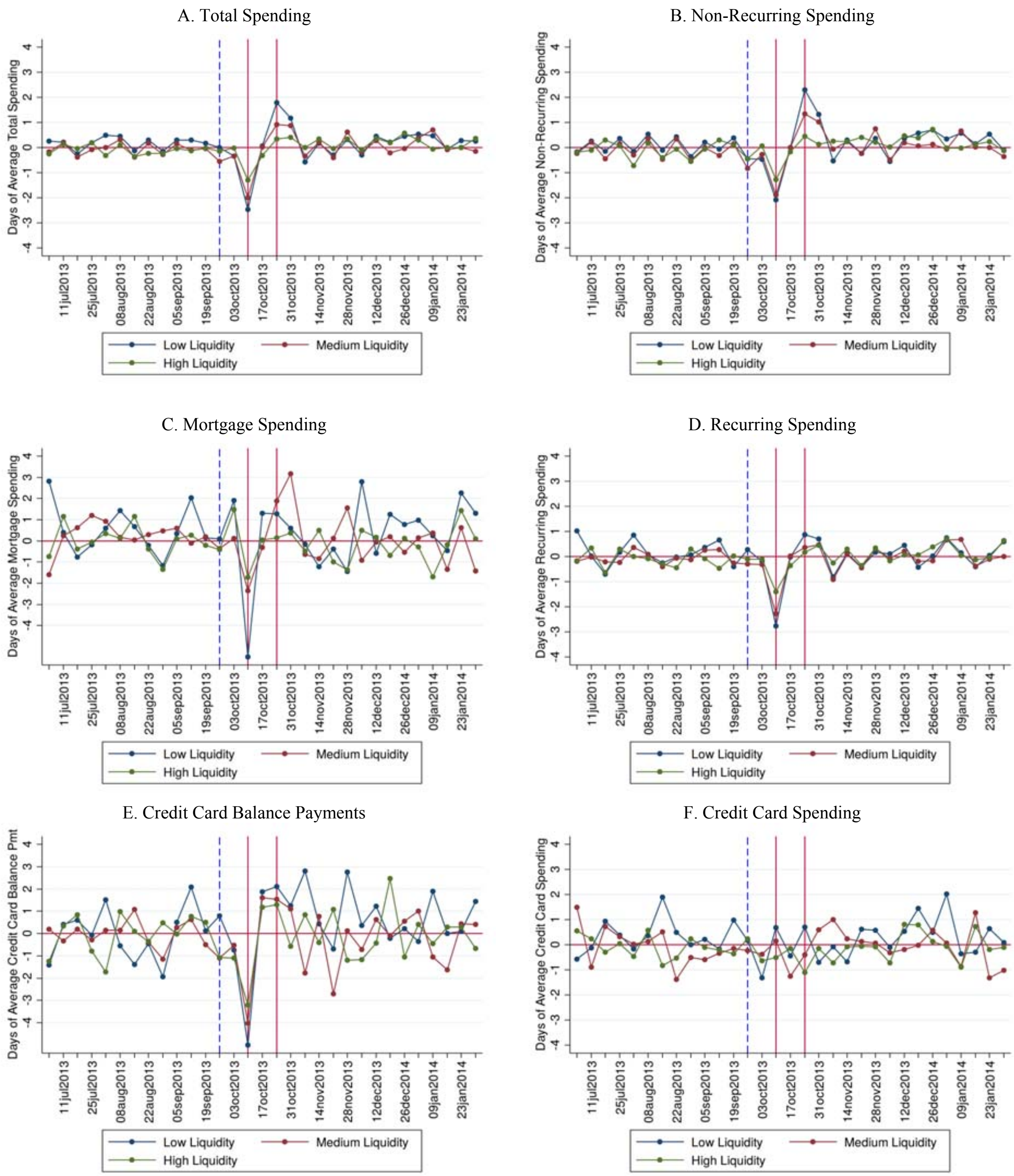


FIGURE 9. ESTIMATED RESPONSE OF SPENDING CATEGORIES TO GOVERNMENT SHUTDOWN BY LIQUIDITY TERCILE

Notes: The spending and payment category in each panel is normalized by the individual-level daily average for that category. The treatment group includes 3,804 individuals and the control group includes 94,669 individuals. Liquidity is expressed as a ratio of checking and savings account balances to average daily spending. Average liquidity is 3, 8, and 54 days for the low, medium, and high liquidity groups respectively.

A. Credit Card Balance, Accounts at Liquidity Risk

B. Credit Card Balance, Accounts not at Risk

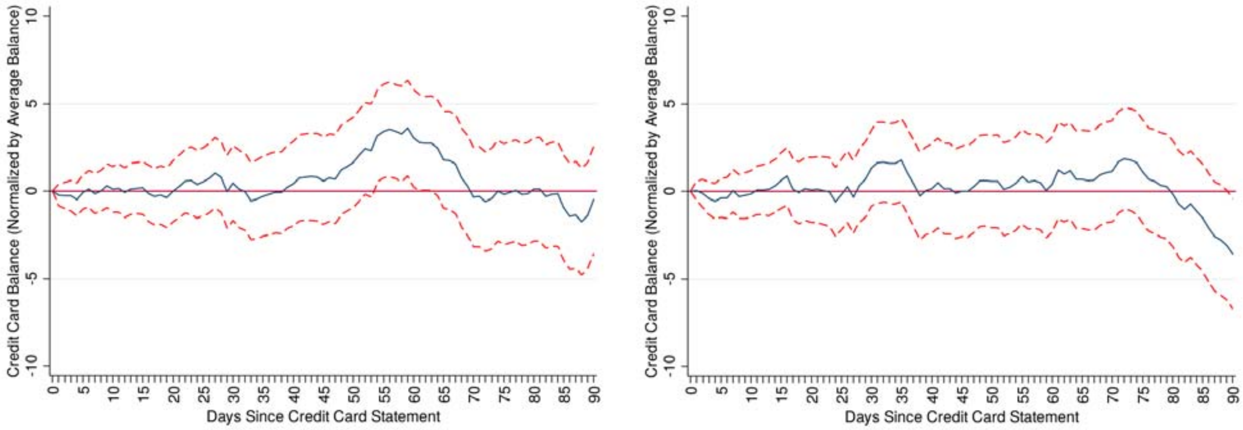


FIGURE 10. ESTIMATED RESPONSE OF CREDIT CARD DEBT TO GOVERNMENT SHUTDOWN

Notes: The sample excludes accounts which never carried revolving credit card debt. Analysis is at the account, not individual level. The figure shows daily account balance normalized by the account-level average balance. Standard errors are clustered at the account level. The horizontal axis is the days since the August 2013 credit card statement. Panel A includes accounts with payment due dates during the pay period affected by the shutdown. Panel B includes accounts with due dates before that pay period. In panel A, the control group observations represent 22,515 individuals, 45,712 accounts, and 4,084,450 days and the treatment group observations represents 1,040 individuals, 2,300 accounts, and 205,746 days. In panel B, the control group observations represent 22,914 individuals, 45,334 accounts, and 4,030,846 days and the treatment group observations represents 999 individuals, 2,203 accounts, and 194,972 days. The outcome variables are winsorized at the upper and lower 2%. Data are winsorized at the 2% level rather than the 1% level in other results to control for the greater outliers in the daily balance data.

TABLE 1—AVERAGE HOURLY COMPENSATION OF FEDERAL EMPLOYEES RELATIVE TO THAT OF PRIVATE-SECTOR EMPLOYEES, BY LEVEL OF EDUCATIONAL ATTAINMENT

	Difference in 2010 Dollars per Hour			Percentage Difference		
	Wages	Benefits	Total Compensation	Wages	Benefits	Total Compensation
High School Diploma or Less	$4	$7	$10	21%	72%	36%
Bachelor's Degree	-	$7	$8	-	46%	15%
Professional Degree or Doctorate	-$15	-	-$16	-23%	-	-18%

Notes: CBO compared average hourly compensation (wages, benefits, and total compensation, converted to 2010 dollars) for federal civilian employees and for private-sector employees with similar observable characteristics that affect compensation—including occupation, years of experience, and size of employer—by the highest level of education that employees achieved. Positive numbers indicate that, on average, wages, benefits, or total compensation for a given education category was higher in the 2005–2010 period for federal employees than for similar private-sector employees. Negative numbers indicate the opposite. Source: Congressional Budget Office based on data from the March Current Population Survey, the Central Personnel Data File, and the National Compensation Survey.

TABLE 2—EMPLOYEE CHARACTERISTICS

	N	Average Weekly Income	Standard Deviation of Weekly Income	Average Weekly Spending	Average Normalized Liquid Balance (days)	Average Credit Card Debt
All Federal Employees	6,792	$1,728	$1,415	$1,855	27	$3,673
Affected by the Shutdown	3,804	$1,727	$1,326	$1,861	26	$3,785
Not affected by the Shutdown	2,988	$1,729	$1,521	$1,849	29	$3,529
Non-Federal Employees	91,692	$1,261	$1,360	$1,362	23	$2,461

Notes: Sample is employees with biweekly paychecks on the same schedule as the government. See text for further details. Normalized Liquid Balance = Average Daily Liquid Balance / Average Daily Spending. The sample is conditional on having accounts that are well linked. Variables are winsorized at the 0.1% upper end. All values are calculated using data from December 2012 to September 2013. Not all users have data for the entire period.

TABLE 3— MPC ESTIMATES

Lag(k)	2	1	0	-1
$\hat{\beta}_{MPC}$	0.0653	0.0213	0.5765	0.0248
	(0.0308)	(0.0335)	(0.0271)	(0.0248)
Observations	98,476	98,476	98,476	98,476
SEE	7.107	6.780	6.602	6.263

Notes: Estimates of equation (2). The right-hand side variable is the change in paycheck in the week starting October 10, 2013 (τ) relative to two weeks earlier. The left-had side variable is weekly spending. Both variables are normalized by the individual-level average daily spending calculated over the entire sample. Separate regressions are estimated for lags and leads of the LHS variable.

Estimation is by instrumental variables with a dummy for an individual being affected by the shutdown as the instrument. Standard errors, corrected for conditional heteroskedacity, are in parentheses.

TABLE 4—LIQUIDITY RATIO

	Treatment Group		Control Group	
Liquidity Ratio Tercile	Mean (Days)	N	Mean (Days)	N
1	2.9	851	2.8	25,105
2	8.4	1131	8.4	24,824
3	54.2	1201	54.7	24,754

Notes: The sample is conditional on having accounts that are well linked. Variables are winsorized at 1%.

Michael Gelman, Shahar Kariv, Dan Silverman, Matthew D. Shapiro, and Steven Tadelis. Harnessing Naturally Occurring Data to Measure the Response of Spending to Income. *Science* 345 (July 11, 2014) 212-215. http://dx.doi.org/10.1126/science.1247727

MICROECONOMICS

Harnessing naturally occurring data to measure the response of spending to income

Michael Gelman,[1] Shachar Kariv,[2] Matthew D. Shapiro,[1,3]* Dan Silverman,[3,4] Steven Tadelis[3,5]

This paper presents a new data infrastructure for measuring economic activity. The infrastructure records transactions and account balances, yielding measurements with scope and accuracy that have little precedent in economics. The data are drawn from a diverse population that overrepresents males and younger adults but contains large numbers of underrepresented groups. The data infrastructure permits evaluation of a benchmark theory in economics that predicts that individuals should use a combination of cash management, saving, and borrowing to make the timing of income irrelevant for the timing of spending. As in previous studies and in contrast to the predictions of the theory, there is a response of spending to the arrival of anticipated income. The data also show, however, that this apparent excess sensitivity of spending results largely from the coincident timing of regular income and regular spending. The remaining excess sensitivity is concentrated among individuals with less liquidity.

Economic researchers and policy-makers have long sought high-quality measures of individual income, spending, and assets from large and heterogeneous samples. For example, when policy-makers consider whether and how to stimulate the economy, they need to know how individuals will react to changes in their income. Will individuals spend differently? Will they save at a different rate or reduce their debt, and when? There are many obstacles to obtaining reliable answers to these important questions. One obstacle is that existing data sources on individual income and spending have substantial limits in terms of accuracy, scope, and frequency.

This paper advances the measurement of income and spending with new high-frequency data derived from the actual transactions and account balances of individuals. It uses these measures to evaluate the predictions of a benchmark economic theory that states that the timing of anticipated income should not matter for spending. Like previous research, it finds that there is a response of spending to the arrival of anticipated income. The data show that, on average, an individual's total spending rises substantially above average daily spending on the day that a paycheck or Social Security check arrives, and remains high for at least the next 4 days. The data also allow the construction of variables that show, however, that this apparent excess sensitivity of spending results in large part from the coincident timing of regular income and regular spending. The remaining excess sensitivity is concentrated among individuals who are likely to be liquidity-constrained.

Traditionally, researchers have used surveys such as the Consumer Expenditure Survey (CEX) to measure individual economic activity. Such surveys are expensive to implement and require considerable effort from participants and are therefore fielded infrequently, with modest-sized samples. Researchers have recently turned to administrative records, which are accurate and can be frequently refreshed, to augment survey research. So far, however, the administrative records have typically represented just a slice

[1]Department of Economics, University of Michigan, Ann Arbor, MI 48109, USA. [2]Department of Economics, University of California, Berkeley, Berkeley, CA 94720, USA. [3]National Bureau of Economic Research (NBER), Cambridge, MA 02138, USA. [4]Department of Economics, Arizona State University, Tempe, AZ 85287, USA. [5]Haas School of Business, University of California, Berkeley, Berkeley, CA 94720, USA.
*Corresponding author. E-mail: shapiro@umich.edu

Table 1. Check versus ACS demographics (percent). The sample size for Check is 59,072, 35,417, 28,057, and 63,745 for gender, age, education, and region, respectively. The sample size for ACS is 2,441,532 for gender, age, and region and 2,158,014 for education.

	Check	ACS
Sex		
Male	59.93	48.59
Female	40.07	51.41
Age		
18–20	0.59	5.72
21–24	5.26	7.36
25–34	37.85	17.48
35–44	30.06	17.03
45–54	15.00	18.39
55–64	7.76	16.06
65+	3.48	17.95
Highest degree		
Less than college	69.95	62.86
College	24.07	26.22
Graduate school	5.98	10.92
Census Bureau region		
Northeast	20.61	17.77
Midwest	14.62	21.45
South	36.66	37.36
West	28.11	23.43

of economic activities. They have not provided simultaneous information about various sources of income and forms of spending.

The data described here result from transactions that are captured in the course of business by Check (https://check.me), a financial aggregation and service application (app). The resulting income data are accurate and comprehensive, in that they capture income from several sources and can be linked to similarly accurate and comprehensive information on spending. These raw data present important technical and conceptual challenges. This paper describes protocols necessary for turning them into a data set with several features that are useful for research and policy analysis.

Check had approximately 1.5 million active users in the United States in 2012. Users can link almost any financial account to the app, including bank accounts, credit cards, utility bills, and more. The application logs into the Web portals for these accounts daily and obtains the user's primary financial data. The data are organized so that users can obtain a comprehensive view of their financial situation.

The data we analyzed are derived from a sample of approximately 75,000 Check users, selected at random from the pool of U.S.-based users who had at least one bank or credit card account, and cover 300 consecutive days spanning 2012 and 2013. The data were de-identified and the analysis was perfomed on normalized and aggregated user-level data as described in the text and supplementary materials (SM). Check does not collect demographic information directly, and instead uses a third party that gathers both publicly and privately provided demographic data, anonymizes them, and provides aggregate tabulations of demographic characteristics of users. Table 1 compares the gender, age, education, and geographic distributions in the Check sample that matched with an email address to the distributions in the U.S. Census Bureau's American Community Survey (ACS), representative of the U.S. population in 2012.

Table 1 shows that the data overrepresent males and those aged 25 to 44. Education levels are broadly similar to those of the U.S. population, and the geographic distribution of Check users is reasonably consistent with that of the U.S. population. Overall, the sample contains large numbers of even the most underrepresented sociodemographic groups. For example, the sample contains about 3000 individuals aged 65 and older. At a point in time, the CEX contains information on approximately 1100 individuals aged 65 and older. We note, however, that the willingness to provide login credentials may select on personal characteristics or increased need for financial organization. The extent of this selection could be assessed with surveys of Check users, the results of which could be compared to those from existing surveys of representative populations. Alternatively, random samples of the population could be encouraged to link their financial accounts to the app, and the transaction and balances of this population could be compared with those of Check users.

Summary statistics for the raw data are provided in tables S1 and S2 of the SM. The data allowed us to calculate total income and to separately identify paychecks and Social Security payments using the description fields of transactions. Similarly, we measured total spending and subcategories of spending. We identified recurring and nonrecurring income and spending by looking for transactions that occurred at regular periodicity and had regular amounts.

We derived two measures of income: The first sums all transactions that represent credits to a user's non–credit card accounts, excluding transfers from one account to another. The second isolates only those transactions that credit paychecks and Social Security payments, using a list of keywords commonly found in the description field. Figure 1 shows the distribution of average monthly income measures at the user level.

Total monthly income depicted in Fig. 1A has a median of $4800 and a mean of $8923. The long and heavy right tail reflects income inequality and also includes large one-off transactions from asset sales. Paycheck and Social Security income, shown in Fig. 1B, is less skewed, with a median of $2900 and a mean of $3951. The figure also displays a kernel density estimate of the distribution of monthly incomes reported in the U.S. Census Bureau's ACS. The income concepts in the ACS and Check data have important differences. Figure 1A shows the distribution of ACS monthly pre-tax household income. The Check data shown in Fig. 1A are net of any (tax) withholding and may be aggregated from either individual or household income. Despite these differences, the ACS and Check distributions are qualitatively similar. Figure 1B shows the ACS distribution of wages, salaries, and the sum of wages and salaries and Social Security payments, which are more closely aligned with their analogs in the Check data. The shape of the ACS distribution is again similar to Check's.

For credit card accounts, we identify spending as transactions that post debits to the account. Non–credit card accounts are similar, but a sum of their debits will overstate spending, because some may represent credit card payments or transfers between accounts. Consequently, spending measures exclude debits we can identify as such payments or transfers either by amount or by transaction description.

We considered three measures of spending: (i) total spending, calculated using the method just described; (ii) nonrecurring spending; and (iii) spending on fast food and coffee shops. See fig. S1 in the SM for the distribution of these average weekly spending measures at the user level. Nonrecurring spending subtracts from total spending both ATM cash withdrawals and expenditures of at least $30 that recur, in the exact same amount (to the cent), at regular frequencies, such as weekly or monthly. It isolates spending that, due to its irregularity, is not easily timed to match the arrival of income. This measure thus uses the amount and timing of spending rather than an a priori categorization based on goods and services, an approach

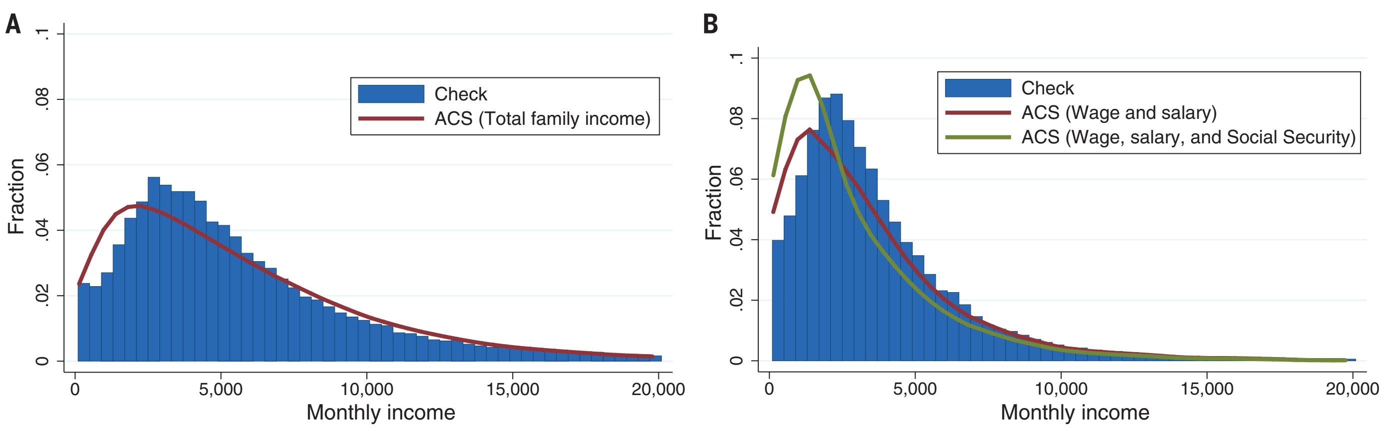


Fig. 1. Distribution of monthly income. (**A**) Total income. (**B**) Paycheck and Social Security income. The figure shows average monthly income across users. Any month in which the user had fewer than 20 days of data was dropped from computation of the average. In (A), the Check distribution represents 61,184 users who have at least one checking or savings account. In (B), both Check and ACS distributions are conditional on having paycheck and Social Security income (47,050 users).

made possible by the distinctive features of the data infrastructure. The fast food and coffee shop measure is identified using keywords from the transaction descriptions. This measure isolates an especially discretionary, nondurable, and highly divisible form of spending, which we used in the analysis of the spending response to anticipated income.

A benchmark theory indicates that the anticipated arrival of a payment should not affect the timing of spending. Specifically, spending should not rise after the arrival of a regular paycheck or Social Security payment. We estimated the excess sensitivity of total, nonrecurring, and coffee shop and fast food spending to the arrival of regular paychecks or Social Security payments. We thus evaluated the possibility that the benchmark theory describes behavior well and that excess sensitivity reflects either the convenience of coordinating recurring expenses with the arrival of regular income, or the intrinsic difficulty of smoothing some forms of spending. We also estimated the excess sensitivity of spending separately for users with different levels of liquidity and different levels of available credit. We thus evaluated the possibility that, as standard enhancements to the benchmark theory indicate, excess sensitivity is a phenomenon of those with inadequate liquidity or credit.

We restricted attention to approximately 23,000 users observed to receive paychecks or Social Security payments at a regular frequency and in regular amounts. A payment is classified as regular in frequency if the median number of days between its arrival is from 13 to 15 or from 26 to 34 and if its coefficient of variation is less than 0.5. The demographic characteristics of users who receive either regular paychecks or regular Social Security payments are remarkably similar to those of the entire sample, as are the distributions of their income, spending, and balances.

Our main econometric specification is

$$x_{ict} = \sum_{j=\text{Mon.}}^{\text{Sun.}} \delta_{jc} + \sum_{k=-7}^{6} \beta_{kc} I_i(\text{Paid}_{t-k}) + \varepsilon_{ict} \qquad (1)$$

where x_{ict} is the ratio of spending of individual i to i's average daily spending in category c, at date t, δ_{jc} is a day-of-week fixed effect, and $I_i(\text{Paid}_{t-k})$ is an indicator equal to 1 if i received a payment at time $t-k$, and equal to 0 otherwise. The β_{kc} coefficients thus measure the fraction by which individual spending in category c deviates from average daily spending in the days surrounding the arrival of a payment. The day-of-week dummies capture within-week patterns of both income and spending.

Figure 2 shows estimates of β_{kc} for the following categories of spending: (A) total, (B) nonrecurring, and (C) coffee shop and fast food spending. The dashed lines are the bounds of the 95% confidence intervals of these estimates based on heteroskedasticity-robust standard errors, with clustering at the individual level. Figure 2A shows that, on average, a user's total spending rises about 70% above its daily average on the day that a regular paycheck or Social Security payment arrives, and remains high for at least the next 4 days.

Total spending includes, however, expenditures such as rent, cable bills, or tuition that are recurring and predictable and whose timing can be adjusted to match the arrival of regular income. Figure 2B shows the excess sensitivity of only nonrecurring spending, confirming that a substantial part (40%) of the excess sensitivity of total spending can be attributed to the convenience of paying major bills automatically

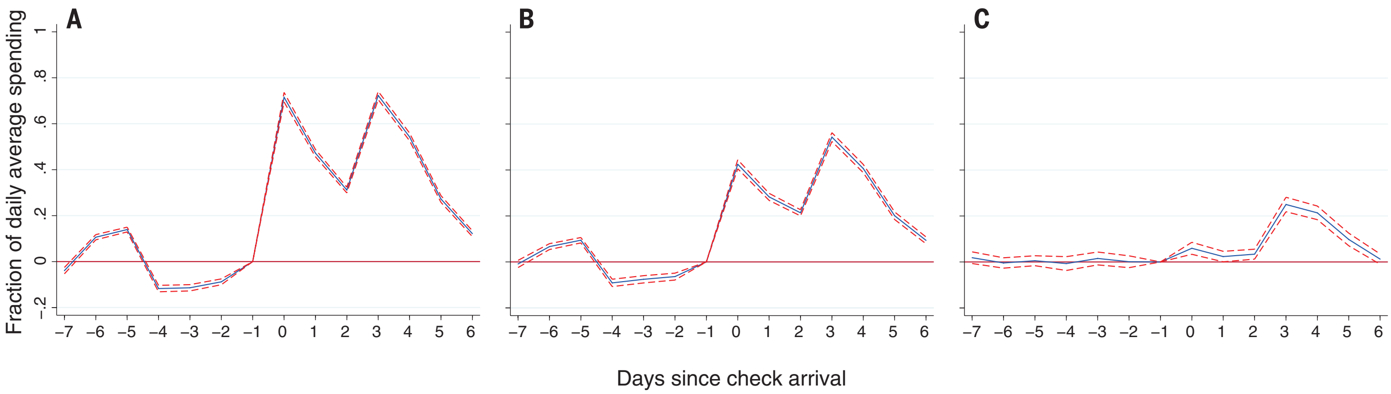


Fig. 2. Response of spending to income: Alternative components of spending. (**A**) Total spending. (**B**) Nonrecurring spending. (**C**) Fast food and coffee shop spending. The solid line represents regression coefficients from Eq. 1. The dashed lines are 95% confidence intervals. Estimates are based on 5,371,244, 5,371,244, and 5,173,594 total observations from 23,985, 23,985, and 23,021 users for panels (A), (B), and (C), respectively.

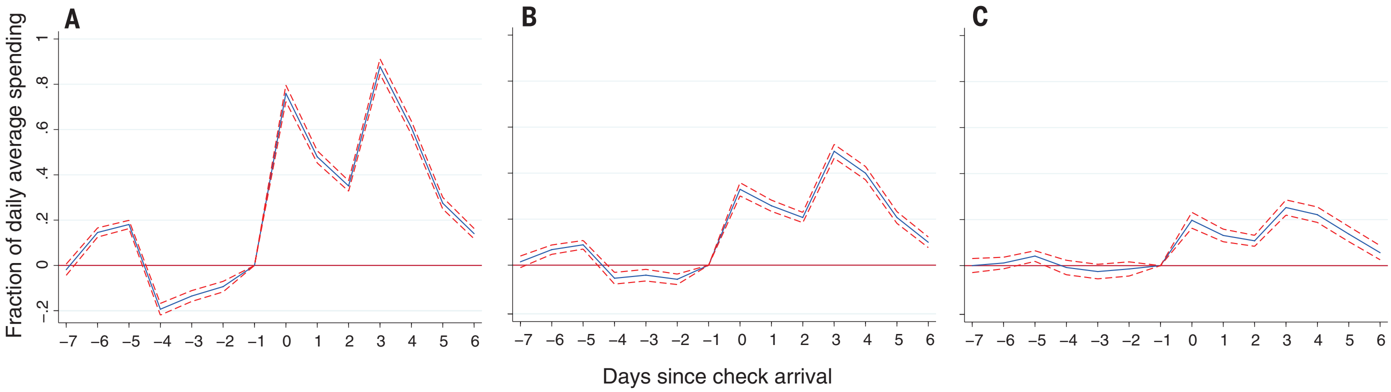


Fig. 3. Response of nonrecurring spending to income: Liquidity ratio. (**A**) Low liquidity. (**B**) Medium liquidity. (**C**) High liquidity. The solid line represents regression coefficients from Eq. 1. The dashed lines are 95% confidence intervals. Estimates are based on 1,784,460, 1,809,839, and 1,769,968 total observations from 7956, 7956, and 7955 users for panels (A), (B), and (C), respectively. The liquidity ratio is defined as the average daily balance of checking and savings accounts normalized by daily average spending.

and avoiding the bad consequences of temporary illiquidity. Given that we defined recurring spending conservatively (i.e., required that it be the same amount to the cent), this estimate is probably a lower bound on how much accounting for it reduces excess sensitivity.

Figure 2C provides still more evidence that the benchmark theory is a better description of behavior than the total spending estimates would suggest. For this imminently divisible and easily smoothed discretionary spending, we observe very modest excess sensitivity to the arrival of predictable income.

We find evidence of individual heterogeneity of excess sensitivity that is consistent with the theory that predicts such behavior among those with insufficient liquidity or available credit, perhaps due to imperfections in credit markets. Figure 3 plots estimates of β_{kc} for nonrecurring spending by terciles of liquidity. We define liquidity for each user as the average daily balance of checking and savings accounts over the entire sample period, normalized by the user's average daily spending. The average user in the lowest tercile has 5 days of spending in cash on hand; the average user in the highest tercile has 159 days. The estimates show that excess sensitivity is significantly more pronounced among those in the lowest tercile of the liquidity distribution.

Figure S2 plots estimates of excess sensitivity by terciles of the available credit utilization distribution. Excess sensitivity is concentrated among users near the limit of their ability to borrow with credit cards. Those who have little liquidity or take their debt levels very close to their limits may be poor at planning or optimizing. The evidence indicates that differences in liquidity and constraints drive heterogeneity of excess sensitivity among Check users.

Many prior studies of spending responses to income have used the CEX quarterly retrospective survey, which records self-reports of income but does not measure its timing precisely. Souleles, for example, uses it to estimate the spending response to the arrival of income tax refunds and overcomes the lack of timing information by calculating from aggregate statistics the likelihood of receiving a refund at various dates (*1*). Parker takes a similar approach and exploits anticipated changes in take-home pay when workers hit the annual cap on the Social Security payroll tax (*2*). Johnson *et al.* and Parker *et al.* measure the timing of some income more precisely by adding special questions to the CEX about tax rebates (*3, 4*).

Some studies use higher-frequency data to estimate spending responses to income. The CEX diary survey records spending daily for 2 weeks but does not collect high-frequency income data. Stephens overcomes this limitation by studying the spending response to the receipt of Social Security benefits, which used to arrive on the same day of each month (*5*). The UK's Family Expenditure Survey collects the most recent paystub of respondents and asks them to track spending for 2 weeks. Stephens uses the paystub to infer the amount and timing of paychecks and estimates the spending response to them (*6*).

These prior studies use a variety of methods, but share an interest in estimating either an elasticity, defined as $\frac{\partial \log(\text{spending})}{\partial \log(\text{income})}$, or a marginal propensity to consume (MPC), defined as $\frac{\partial(\text{spending})}{\partial(\text{income})}$. Table S3 summarizes the key features of these prior estimates and compares them to analogous aspects of our study.

The studies differ in the time frame over which they measure spending changes in response to a change in income. This makes the levels of their estimated elasticities or MPCs difficult to compare. For our study, we present the point estimate of effects on the first day after the income arrives; that is β_{1c} from Eq. 1 for the elasticity of spending in category c. For the MPC we present the γ_{1c} from the equation

$$x_{ict} = \alpha_{ic} + \sum_{j=\text{Mon.}}^{\text{Sun.}} \delta_{jc} + \sum_{k=-7}^{6} \gamma_{kc} \text{Payment}_{ic,t-k} + \varepsilon_{ict} \tag{2}$$

where x_{ict} is the ratio of spending of individual i to i's average daily spending in category c, at date t; δ_{jc} is a day-of-week fixed effect; α_{ic} is a user fixed effect; and $\text{Payment}_{ic,t-k}$ is the ratio of the amount of the payment received by individual i divided by i's average daily spending in category c, at date $t-k$. Analogously, table S3 presents only the shortest-run effects reported in all the other studies. Although our and other studies estimate larger impacts at longer horizons, the central conclusion of table S3 about the relative precision of the estimates is not affected by the choice of horizon.

The prior estimates are important and influential but, as table S3 shows, they often lack precision. Among studies of the quarterly CEX data, Hsieh is unusual in its precision (*7*). The last four rows of table S3 include the confidence intervals for our estimates of both the elasticity and the MPC. These intervals are small, both economically and relative to other studies. Only Broda and Parker provides estimates that are as precise as those from the Check data (*8*). That paper uses Homescan data and estimates precisely an MPC out of tax rebates near 0. These estimates rely on surveys to determine receipt of the rebate, however, and would be attenuated if those reports are subject to error. The Homescan spending data are also limited in scope, largely capturing only goods with Universal Product Codes. Moreover, the Check data allow estimates of the response to routine payments such as paychecks and Social Security payments, not just particular payments such as tax rebates.

Related studies of administrative data also provide accurate measures of spending but do not cover it comprehensively. For example, Agarwal *et al.* use data from a single credit card company to study the spending response to tax rebates; they can thus track the effects of the rebate on a single account but not on overall spending (*9*). Kuchler makes use of more comprehensive administrative data collected from a debt management Web site, but the number of users (556) is relatively small (*10*). The financial application Mint (https://www.mint.com/) has a complementary data infrastructure that it is using to construct monthly time series of spending by types of goods (*11*). It has not been used for research along the lines of the estimates in this paper.

In policy discussions before the 2008 tax rebates, the Congressional Budget Office and others cited the point estimates of the effect of the 2001 rebate from Parker, Johnson, and Souleles, but not the substantial uncertainty about that estimate documented in that paper and in table S3 (*3, 12*). More generally, estimates of spending rates from different changes in income play a key role in the evaluation of the American Recovery and Reinvestment Act (*13*), making the stakes in getting credible and precise estimates of these parameters very high. This paper shows how economic theory and policy can benefit from analysis made possible with naturally occurring data such as those provided by Check.

REFERENCES AND NOTES

1. N. S. Souleles, *Am. Econ. Rev.* **89**, 947–958 (1999).
2. J. A. Parker, *Am. Econ. Rev.* **89**, 959–973 (1999).
3. D. S. Johnson, J. A. Parker, N. S. Souleles, *Am. Econ. Rev.* **96**, 1589–1610 (2006).
4. J. A. Parker, N. S. Souleles, D. S. Johnson, R. McClelland, *Am. Econ. Rev.* **103**, 2530–2553 (2013).
5. M. Stephens Jr., *Am. Econ. Rev.* **93**, 406–422 (2003).
6. M. Stephens Jr., *Econ. J.* **116**, 680–701 (2006).
7. C.-T. Hsieh, *Am. Econ. Rev.* **93**, 397–405 (2003).
8. C. Broda, J. A. Parker, "The economic stimulus payments of 2008 and the aggregate demand for consumption" (manuscript, Sloan School of Management, Cambridge, MA, 2014).
9. S. Agarwal, C. Liu, N. S. Souleles, *J. Polit. Econ.* **115**, 986–1019 (2007).
10. T. Kuchler, "Sticking to your plan: Hyperbolic discounting and credit card debt paydown" (manuscript, Stern School of Business, New York, 2013).
11. Intuit, Intuit consumer spending index (Intuit, Mountain View, CA, 2013).
12. Congressional Budget Office, *Options for Responding to Short-Term Economic Weakness* (U.S. Congressional Budget Office, Washington, DC, 2008).
13. Congressional Budget Office, *Estimated Impact of the American Recovery and Reinvestment Act on Employment and Economic Output from July 2010 Through September 2010* (U.S. Congressional Budget Office, Washington, DC, 2010).

ACKNOWLEDGMENTS

This research was supported by a grant from the Alfred P. Sloan Foundation. M.D.S. acknowledges additional support through the Michigan node of the NSF-Census Research Network (NSF grant SES 1131500). This paper has benefited from suggestions by the participants of the NBER Summer Institute, the Conference on Economic Decisionmaking (Aspen, Colorado), and several seminars. A data set for replicating the results of this paper is available through the University of California Berkeley Econometrics Lab (EML) at https://eml.berkeley.edu/cgi-bin/HarnessingDataScience2014.cgi. To access the data, users must register with EML and agree to terms of use. The data set contains no personal or account identifiers. The data are aggregated and transformed so that they reveal no personally identifying information.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/345/6193/212/suppl/DC1
Materials and Methods
Figs. S1 and S2
Tables S1 to S3
References

28 October 2013; accepted 13 June 2014
10.1126/science.1247727

Commentary

Big Data & Society
April–June 2014: 1–6
© The Author(s) 2014
DOI: 10.1177/2053951714538417
bds.sagepub.com

Official statistics and Big Data

Peter Struijs, Barteld Braaksma and Piet JH Daas

Abstract

The rise of Big Data changes the context in which organisations producing official statistics operate. Big Data provides opportunities, but in order to make optimal use of Big Data, a number of challenges have to be addressed. This stimulates increased collaboration between National Statistical Institutes, Big Data holders, businesses and universities. In time, this may lead to a shift in the role of statistical institutes in the provision of high-quality and impartial statistical information to society. In this paper, the changes in context, the opportunities, the challenges and the way to collaborate are addressed. The collaboration between the various stakeholders will involve each partner building on and contributing different strengths. For national statistical offices, traditional strengths include, on the one hand, the ability to collect data and combine data sources with statistical products and, on the other hand, their focus on quality, transparency and sound methodology. In the Big Data era of competing and multiplying data sources, they continue to have a unique knowledge of official statistical production methods. And their impartiality and respect for privacy as enshrined in law uniquely position them as a trusted third party. Based on this, they may advise on the quality and validity of information of various sources. By thus positioning themselves, they will be able to play their role as key information providers in a changing society.

Keywords

Big Data, official statistics, European Statistical System

Introduction

The advent of Big Data is expected to have a big impact on organisations for which the production and analysis of data and information is core business. National Statistical Institutes (NSIs) are such organisations. They are responsible for official statistics, which are heavily used by policy-makers and other important players in society. Arguably, the way NSIs take up Big Data will eventually have implications for all of society.

Official statistics play a key role in modern society. NSIs aim at providing information on all important aspects of society in an impartial way, and according to the highest scientific standards. Information that fulfils these demands is used in public discussion, forms the basis of policy decisions, is required for business use, feeds scientific research, is used in education and so on. Official statistics can only meet this demand if they can be trusted. In advanced societies, official statistics are often taken for granted, but where trust is lacking, society misses an important pillar for informed discussion and evidence-based policy-making.

Professional standards play a vital role in securing trust in official statistics. Statisticians have their own ethics code (United Nations, 2013), which includes an absolute respect for the confidentiality of data provided by respondents. Data collected for statistical purposes may never be disclosed and may never be used for other purposes. At the level of the European Union (EU), quality norms have been codified in the so-called Statistics Code of Practice (Eurostat, 2014). The trust earned by respecting professional standards is also the basis for a privileged position of NSIs in respect of data acquisition. Many NSIs have access by law to government data sources and have the power to collect data from other parties, often without having to pay the provider. Moreover, for statistical purposes, many NSIs are allowed to link data from different sources.

Statistics Netherlands, Heerlen, The Netherlands

Corresponding author:
Peter Struijs, Statistics Netherlands, P.O. Box 4481, Heerlen, 6401 CZ, The Netherlands.
Email: p.struijs@cbs.nl

Given this role for NSIs, what does the emergence of Big Data mean for official statistics? This question is addressed in this contribution, but as we will see, there are many reasons why the role of NSIs in the Big Data era is not 'given'. In order to keep a sound and trusted basis of information for society to rely on, we argue that NSIs may have to adapt to the changing context in which they operate.

Official statistics in a changing context

In respect of information, society is changing rapidly. For example, there is an enormous growth of data that is gathered and recorded in myriad ways: from satellite and sensory data, to social network and transactional data and so on. The availability of data is also expanding and becoming the foundation of business models. Information is becoming more visual and interactive. Information and communication technology is becoming ever more advanced, processing power and data storage capacity is continuously rising, cloud solutions are emerging and applications are becoming more intelligent. These developments have been described in more depth and detail by many observers, such as Mayer-Schönberger and Cukier (2013).

These changes have many impacts on societies. For one, the increased gathering of data and the commercial and social possibilities of data usage influence public opinion on privacy. Some are concerned if their data are re-used without their consent, for commercial reasons or otherwise. Others do not mind so much, if this means that services are provided for free. Many people voluntarily share information on social networks without caring for privacy. People have less patience to fill in questionnaires, especially if the data requested have been registered somewhere else already. Government agencies are expected to be more forthcoming in providing data. Governments have reacted to the changes by formulating policies on, for instance, open data and availability of public sector information, also at the EU level (European Union, 2013).

How have NSIs responded? Until around the 1980s, data were essentially a scarce commodity with a high price. Before the era of Big Data, information was not readily available but had to be collected for a particular purpose. Official statistical information based on survey data had a unique value: there simply was no alternative. For example, population census data, collected door to door, was immensely valuable to policy-makers, researchers and other users. In the last few decades, data collected by public administrations have become increasingly accessible for statistical purposes, stimulated in part by IT developments. Statistical data collection by means of questionnaires was supplemented and increasingly replaced by administrative data sources. Nowadays, some countries do not conduct extensive population surveys anymore but compile census statistics by combining and analysing data from several administrative sources. NSIs became more integrated in the information architecture of the government. In this way, the burden on persons and businesses to respond to questionnaires was considerably reduced.

In the context of all of these developments, the information provided by NSIs still remained unique. In particular, the possibility of combining data from different sources made official statistics even more valuable, since in many countries no other organisation was positioned to do so. In parallel, efforts also increased to standardise and harmonise these various sources of official statistics, especially in the EU. Supported by legislation, official statistics in the EU are now considered a system, the so-called European Statistical System, or ESS.[1]

However, Big Data is changing the environment of the NSIs once more as data scarcity is becoming less of an issue. For NSIs, there are potential benefits as new data sources and opportunities emerge. But it also makes the products of NSIs potentially less unique, since other players in the information market may start – and have actually started – producing statistics, for instance, on inflation, such as the Billion Prices Project of MIT.[2]

Let us first look at the opportunities for NSIs offered by Big Data. There is a huge potential for new statistics (Daas et al., 2013). Location data for mobile phones could be used for almost instantaneous daytime population and tourism statistics (De Jonge et al., 2012). Social media messages could be used for several types of indicators, such as an early indicator of consumer confidence. Inflation figures could be derived from price information on the web, and so on. In addition, Big Data sources may be used to substitute or supplement more traditional data sources, such as questionnaire and administrative data. For instance, data collection by questionnaire on road use may not be necessary anymore if detailed traffic loop data, i.e. data from sensors in roads, become available (Struijs and Daas, 2013).

However, in order to realise these opportunities, a number of challenges have to be overcome, which are generally applicable to all uses of Big Data as an information source and as such are not unique to NSIs.

Challenges and issues

Some of the biggest challenges that statisticians face in their use of Big Data concern methodology. Many Big Data sources, such as social media messages, are composed of observational data and are not deliberately designed for data analysis, and thus do not have a well-defined target population, structure and quality. This makes it difficult to apply traditional statistical methods, based on sampling theory (Daas and Puts, 2014a). The unstructured nature of many Big Data sources makes it even more difficult to extract meaningful statistical information. For many Big Data sources, the interpretation of the data and its relationship with social phenomena of interest is far from obvious. For example, public Facebook messages in the Netherlands clearly reflect general sentiment in some sense, but it is far from clear how exactly (Daas and Puts, 2014b). Moreover, if such data are to be used as a source for a population sentiment indicator, one would like to know the relationship between the population of persons writing public messages on Facebook and the population at large. This is challenging without falling back to surveys. Furthermore, the population of persons using social media is likely to change over time, making a comparison to the population at large even more challenging.

For NSIs, a key question concerns how the quality of official statistics can be guaranteed if they are based on Big Data. To address this, new methodologies and forms of interpretation need to be developed. Take for example mobile phones. If data from mobile phone providers are used for statistics on, say, population mobility, the statistician has to interpret anonymised detailed call records from individual phones and derive information about the behaviour of the people using them. That means dealing with the fact that measureable phone activity may vary during the day, some persons may have multiple mobile phones or none, children carry mobile phones which are registered to their parents, phones may be switched off, etc. For social media, even more questions arise such as who is the author of a message. While some methodological remedies have already been developed to some extent, such as deriving the gender and age of a social media user by the known correlation between sex, age and choice of words, these still pose a challenge, as explained above.

Privacy and legal issues form another challenge. The prevention of the disclosure of the identity of individuals is an imperative, but this is difficult to guarantee when dealing with Big Data. Since legislation typically lags behind the emergence of new social phenomena, the legal situation for cases involving Big Data is not always clear. In such cases, one may have to fall back on ethical standards to decide on whether and how to use Big Data. Other legal issues relate to copyright and the ownership of data. Even if data may legally be used, this does not imply that it is wise or appropriate to do so. Of critical importance is the implication of any use of Big Data for the public perception of an NSI as this has a direct impact on trust in official statistics. These concerns have been heightened by the revelations that intelligence agencies are among the most active Big Data users. For NSIs, it is critical that these concerns be addressed through practices such as being transparent about what and how Big Data sources are used. Other mechanisms could also be developed. For example, in some cases it might be feasible to adopt informed consent approaches. Some mobile phone subscription contracts, for instance, offer an opt out to the subscriber for using their data for other purposes than providing the phone service. If the opt out rate is not too high, this does not seriously affect the usability of mobile phone data for statistical purposes.

Another obvious challenge is the processing, storage and transfer of large data sets. Technological advances like increases in computing power, larger storage facilities and high bandwidth data channels may partly solve these issues. Having data processed at the source, thus preventing the transfer of large data sets and the duplication of storage, may also be considered. These technological challenges include mechanisms for ensuring the security of data, which is of the utmost importance because of privacy and confidentiality concerns and makes, for example, cheap cloud-based solutions less attractive.

Another issue is the possible volatility of Big Data sources, given the fact that official statistics often take the form of time series analyses. For many users, the continuity of these series is of the utmost importance. Still another issue is the skills required for dealing with Big Data. Modern data scientists may be better equipped than traditionally trained statisticians. Probably more important is the need for a different mind-set as the use of Big Data may imply a paradigm shift, including an increased and modified use of modelling techniques (Daas and Puts, 2014a; Struijs and Daas, 2013).

Collaboration

Faced with these challenges, NSIs have recognised the necessity of not working in isolation but collaborating with each other and others outside the community of official statistics. This collaboration is often exploratory and may be aimed at sharing knowledge and experiences, but there are already examples of collaboration that go further.

From the perspective of NSIs, several types of partners are of interest. First of all, the potential providers of Big Data are essential partners: if they do not grant access to their data, the story is over before it starts. Data owners have their own concerns and, like NSIs, they are subject to privacy rules. This may complicate collaboration even if they have a positive outlook and approach. But since Big Data sources are not designed for statistical use, such collaboration is also essential in order to obtain good knowledge of the provenance of such sources. Additionally, for statistical production, it may be more efficient to have data processed at the site of collection and storage. In such cases, the assumption that data can be provided for free may no longer hold. On the other hand, statisticians also have much to offer such as providing analytic insights that may help data owners understand their data better. Doing complex statistical analyses is core business for NSIs, but not for, say, a mobile phone company. In these and other ways, the relationship with data providers could potentially become true partnerships. For example, one specific role that NSIs could play is that of a trusted third party. In a competitive market, competitors will be reluctant to share sensitive data among each other. But they might be willing to share it with an NSI who compiles statistical information that is beneficial to all.

Collaboration between NSIs and academia may grow as well. Universities have historically been natural partners for NSIs. It stands to reason that such collaboration will extend to the field of Big Data, for instance, in solving methodological problems, developing technical solutions and training future data scientists. Such collaboration is also being supported by public funders who are facilitating research and innovation partnerships through targeted grants.[3] By working in partnership, researchers in universities and NSIs could better leverage such opportunities.

Furthermore, there are many commercial partners with which NSIs could collaborate. Google and Facebook are two examples for which Big Data forms the core of their business model. Their knowledge and the data to which they have access may be very relevant to NSIs. IT companies also possess relevant knowledge on Big Data processing and storage, security, cloud processing, etc. Apart from the provision of paid services, collaboration may be of interest to them with a view to obtaining statistical expertise and for benchmarking or validating their information products.

Collaboration between NSIs in the field of Big Data has already started. Big Data has become a prominent subject at many statistical meetings and conferences in Europe, such as the 2013 New Techniques and Technologies for Statistics (NTTS) conference,[4] a scientific conference organised by Eurostat, and the 'ESS Big Data event 2014' in Rome.[5] The directors-general of all European NSIs met in Scheveningen in September 2013 to learn about Big Data and adopted the Scheveningen Memorandum (DGINS, 2013). This memorandum calls for an international strategic approach to Big Data and plans for the adoption of an action plan and roadmap by mid-2014.

For some time already, Big Data has been an important topic for the UNECE, the United Nations Economic Commission for Europe. Collaboration at that level resulted in an overview paper about the implications of Big Data for official statistics (UNECE, 2013a). Seminars have been held, facilitating the exchange of knowledge, for instance, on statistical data collection.[6] In 2014, the UNECE went one step further in facilitating cross-national work through a project with the following stated objectives:

a. to identify, examine and provide guidance for statistical organisations to act upon the main strategic and methodological issues that Big Data poses for the official statistics industry;
b. to demonstrate the feasibility of efficient production of both novel products and 'mainstream' official statistics using Big Data sources, and the possibility to replicate these approaches across different national contexts;
c. to facilitate the sharing across organisations of knowledge, expertise, tools and methods for the production of statistics using Big Data sources (UNECE, 2013b).

The future of official statistics

What does the advent of Big Data mean for official statistics? As we have argued, it provides many opportunities. But in order to make optimal use of Big Data, a number of issues have to be addressed. This calls for increased collaboration with private and academic partners who have access to specific Big Data sources and knowledge, but also between NSIs. The relationship between the various stakeholders will involve each partner building on and contributing different strengths and will likely result in flexible networks. Such networks are flexible in the sense that membership of the network and the contribution of partners depend on actual needs instead of being fixed in advance for a long time.

Seen from the viewpoint of NSIs, there are also potential risks. Official statistics are facing more competition. In a time of growing data abundance, generating statistical information that is potentially relevant to society is no longer an activity intrinsically restricted to NSIs. And even the traditional advantage of NSIs, being legally allowed to collect data and combine data sources, is eroding. It may not be possible to combine survey data and administrative data with Big Data sources at the micro-level, which reduces the relative disadvantage traditionally faced by the competition.

For some statistics, Big Data sources cannot be easily envisaged as alternatives to more traditional data sources. This certainly holds for official figures on government finance and economic growth, which are heavily used for decision-making at both the national and international level. But, given the increasing competition that data generated by other sources is presenting to the role of NSIs as bearers of official statistics, a strategic reassessment is needed. This could include fundamental questions such as whether statistics based on Big Data sources should be a core activity of NSIs, or if some data and information should be provided by other market actors, or if NSIs can or should provide new services in this context.

But by posing these questions, we return to the basic premise that society's access to impartial statistical information must be maintained at all times, either by NSIs or other parties. In choosing a position, NSIs could build on and promote their strengths and unique position. Especially at a time of competing and multiplying data sources, their impartiality and respect for privacy as enshrined in law uniquely position them as a trusted third party. They also have unique knowledge of official statistical production methods. Finally, they continue to have privileged access to government data sources that provide unique information and knowledge and have the authority to collect data for statistical purposes that because of privacy considerations will never be available to businesses.

As a consequence, in the context of the challenges of Big Data sources, NSIs will remain important providers of official statistics. And where other organisations are able to provide statistical information to the public, rather than competing, NSIs could build on their position as an impartial, trusted third party and their expertise to advise on the quality and validity of information of these various sources. Possibly, then, providers of Big Data may even seek validation of their data from NSIs, thereby opening up yet another possibility for new partnerships.

The future of official statistics in the age of Big Data is still a matter of some deliberation and experimentation. But what is clear already is that the international statistical community needs to adapt to a new reality and respond to the opportunities and challenges it provides. To do so calls for greater collaboration with players inside and outside the statistical community, through the formation of flexible networks that can forge new ways of generating statistical data. For all engaged with statistics, we think the Big Data era is a most exciting time.

Acknowledgements

The views expressed in this contribution are those of the authors and do not necessarily reflect the position of Statistics Netherlands. The authors wish to thank the editors for their valuable suggestions for improvements.

Declaration of conflicting interest

The authors declare that there is no conflict of interest.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Notes

1. http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/ess/ess_news
2. http://bpp.mit.edu/
3. The current EU framework programme for research and innovation, Horizon 2020, is an example (European Commission, 2013), which mentions Big Data specifically.
4. http://www.cros-portal.eu/content/ntts-2013
5. http://www.cros-portal.eu/content/big-data-event-2014
6. http://www.unece.org/stats/documents/2013.09.coll.html

References

Daas PJH and Puts MJH (2014a) Big Data as a source of statistical information. *The Survey Statistician* 69: 22–31. Available at: http://isi.cbs.nl/iass/N69.pdf (accessed 22 May 2014).

Daas PJH and Puts MJH (2014b) Social media sentiment and consumer confidence. *Paper for the workshop on using Big Data for forecasting and statistics*, Frankfurt, Germany, 7–8 April. Available at: http://www.ecb.europa.eu/events/pdf/conferences/140407/Daas_Puts_Sociale_media_cons_conf_Stat_Neth.pdf?409d61b733fc259971ee5beec7cedc61 (accessed 22 May 2014).

Daas PJH, Puts MJ, Buelens B, et al. (2013) Big Data and official statistics. *Paper for the 2013 NTTS conference*, Brussels, Belgium, 5–7 March. Available at: http://www.cros-portal.eu/sites/default/files/NTTS2013fullPaper_76.pdf (accessed 22 May 2014).

De Jonge E, Van Pelt M and Roos M (2012) *Time patterns, geospatial clustering and mobility statistics based on mobile phone network data*. Discussion paper 201214, Statistics Netherlands. Available at: http://www.cbs.nl/NR/rdonlyres/010F11EC-AF2F-4138-8201-2583D461D2B6/0/201214x10pub.pdf (accessed 22 May 2014).

DGINS (2013) Scheveningen memorandum on Big Data and official statistics. Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/0_DOCS/estat/SCHEVENINGEN_MEMORANDUM%20Final%20version.pdf (accessed 22 May 2014).

European Commission (2013) Horizon 2020, the EU framework programme for research and innovation. Available at: http://ec.europa.eu/programmes/horizon2020/ (accessed 22 May 2014).

European Union (2013) Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013, amending Directive 2003/98/EC on the re-use of public sector information. Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:175:0001:0008:EN:PDF (accessed 22 May 2014).

Eurostat (2014) European statistics code of practice. Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/code_of_practice (accessed 22 May 2014).

Mayer-Schönberger V and Cukier K (2013) *Big Data: A Revolution that Will Transform How We Live, Work, and Think*. London: John Murray Publishers.

Struijs P and Daas PJH (2013) Big Data, big impact? *Paper presented at the seminar on statistical data collection*, Geneva, Switzerland, 25–27 September 2013. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2013/mgt1/WP31.pdf (accessed 22 May 2014).

UNECE (2013a) What does 'Big Data' mean for official statistics? *Paper prepared on behalf of the high-level group for the modernisation of statistical production and services*, 10 March. Available at: http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170614 (accessed 22 May 2014).

UNECE (2013b) The role of Big Data in the modernisation of statistical production. Project plan. Available at: http://www1.unece.org/stat/platform/display/msis/Final+project+proposal%3A+The+Role+of+Big+Data+in+the+Modernisation+of+Statistical+Production (accessed 22 May 2014).

United Nations (2013) Fundamental principles of official statistics. Available at: http://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf (accessed 22 May 2014).

Finding errors in Big Data

Weeding out mistakes hidden among billions of data points seems an impossible task. **Marco Puts**, **Piet Daas** and **Ton de Waal** put forward some solutions

No data source is perfect. Mistakes inevitably creep in. Spotting errors is hard enough when dealing with survey responses from several thousand people, but the difficulty is multiplied hugely when that mysterious beast Big Data comes into play.

Statistics Netherlands is about to publish its first figures based on Big Data – specifically road sensor data, which counts the number of cars passing a particular point. Later, we plan to use cell phone data for statistics on the daytime population and tourism, and we are considering an indicator to capture the "mood of the nation" based on sentiment expressed through social media.[1]

Statistics derived from unedited data sets of any size would be biased or inaccurate. But the challenge Statistics Netherlands faces in dealing with Big Data sets is to find data editing processes that scale up appropriately to allow quick and efficient cleaning of a huge number of records.

How huge? For the sentiment indicator, we plan to use 3 billion public messages predominantly gathered from Facebook and Twitter,[2] and for the road sensor data there are 105 billion records. But size is not the only distinguishing characteristic of a Big Data set.

A clear, generally accepted definition of "Big Data" does not exist, though descriptions often refer to the three Vs: volume, velocity, and variety.[3] So, not only do we have a large amount of data to deal with (volume), but the frequency of observations is very high (velocity). For the road sensor data, for example, we have data on a minute-by-minute basis. Big Data also tends to be "messy" in comparison to traditional data (variety). Again, for the road sensor data, we only know how many vehicles passed by. We do not know who drove the cars. In addition, background characteristics, which are important for data editing and estimation methods, are lacking, thus making such methods difficult to apply.

A big problem

Our experience with cleaning large data sets started a few years before we began to study the use of Big Data for statistical purposes. In those days we were investigating how to edit and impute large amounts of administrative data. Administrative data can be high-volume, but differ from Big Data with respect to velocity and variety. We learnt that finding errors in large administrative data sets is already a challenge. Automatic editing techniques and graphical macro-editing techniques (see box, page 28) work best for such data sets.

In order to apply graphical macro-editing to large administrative data sets we applied and (further) developed visualisations. An example of such a visualisation is the "tableplot". A tableplot can be applied in two ways: to detect implausible or incorrect values, or to monitor the effects of the editing process on the quality of the data. In a tableplot, a quantitative variable is used to order the data for all variables shown. The ordered records are divided into a certain number of equally sized bins. For each bin, the mean value is calculated for numerical variables, and category fractions are determined for categorical variables, where missing values are considered as a separate category. These results are subsequently plotted. A disruptive change in the distribution in a tableplot can indicate the presence of errors. Moreover, a non-uniform distribution over the columns can indicate selectivity. Finally, the distribution of correlated variables can be examined by looking at the value distribution in the unsorted columns.

Figures 1 and 2 show tableplots for the Dutch annual Structural Business Statistics (SBS), based on unedited and edited data, respectively. These relatively small data sets – in comparison to Big Data, that is – are used to illustrate the benefits of applying visualisation methods for monitoring the

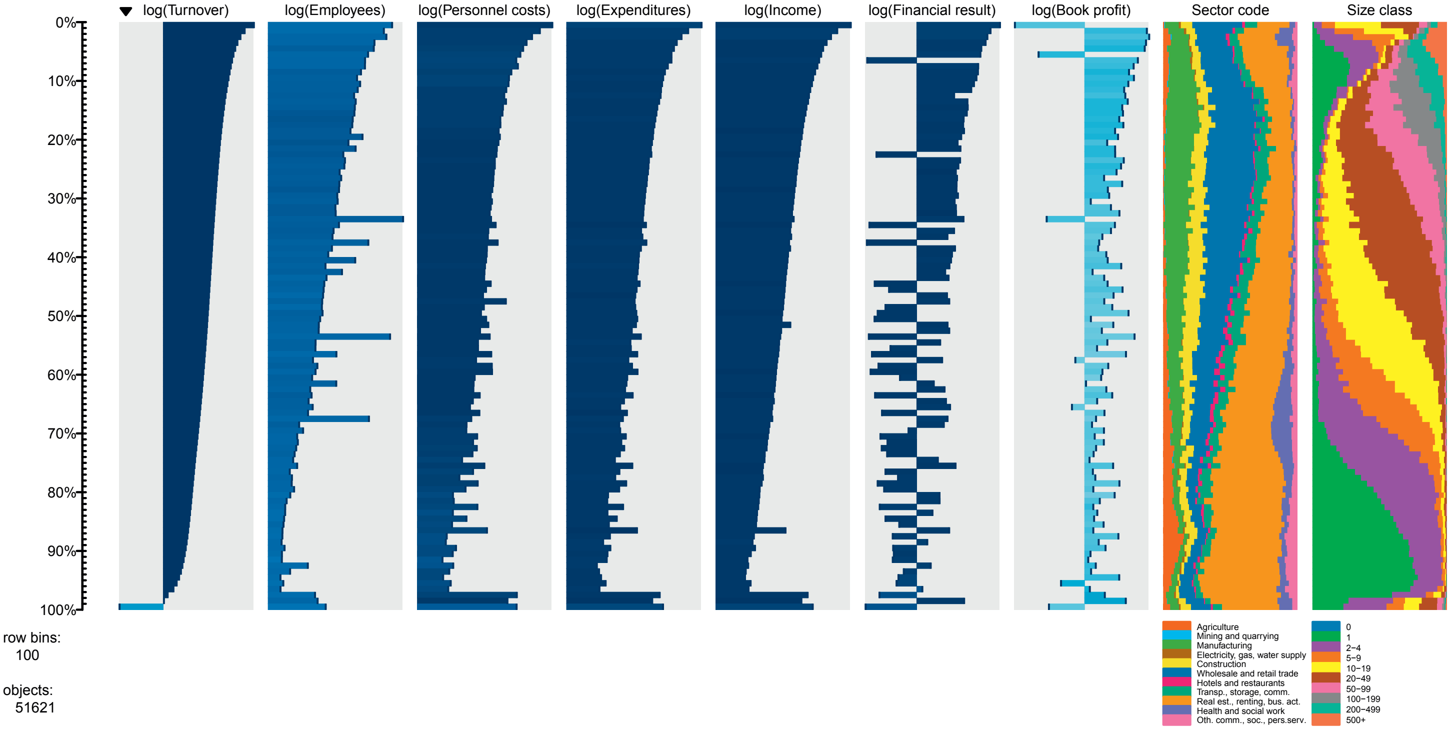


Figure 1. Tableplot of unprocessed SBS data. When sorted on turnover (left-most column) a considerable number of the other numeric variables display a clear – and predictable – downward trend occasionally distorted with large values[4]

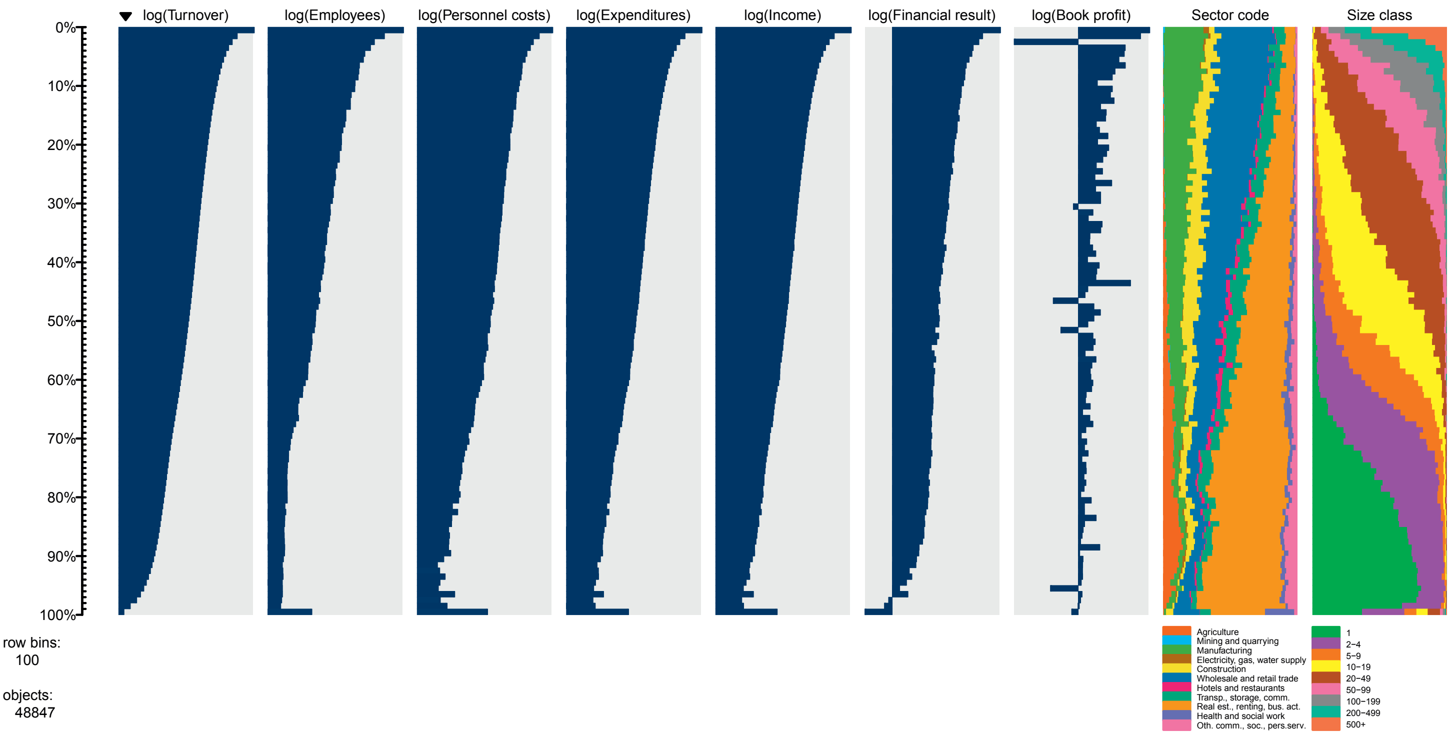


Figure 2. Tableplot of edited SBS data. After data editing and sorting on turnover, the majority of the quality issues have been solved as indicated by the smooth distribution of the variables shown[4]

editing process. The SBS survey covers the economic sectors of industry, trade, and services. Survey data is received from approximately 52 000 respondents annually. Topics covered in the questionnaire include turnover, number of employed persons, total purchases, and financial results. Figure 1 was created by sorting on the first column, "turnover", and dividing the 51 621 observed units into 100 bins, so that each row bin contains approximately 516 records. A subset of approximately 49 000 records was

Cleaning survey data: a small data perspective

In the (distant) past, *manual editing* was used with the intention of correcting all data in every detail. Data was checked and adjusted in separate steps. The editing process thus consisted of cycles where records often had to be examined and adjusted several times, which made for a time-consuming and costly process.

Interactive editing is also a manual activity, where, in principle, all records are examined, and if necessary, corrected. The difference with respect to manual editing is that the effects of adjusting the data can be seen immediately on a computer screen. This immediate feedback directs one to potential errors in the data and enables one to examine and correct each record only once. Interactive editing typically uses edit rules, that is, rules capturing the subject-matter knowledge of admissible (combinations of) values in each record – a male cannot be pregnant, for example – to guide the editing process.

Efficiency is further increased by *selective editing*: identify the records with potentially influential errors and restrict interactive editing to those records only. The most common form of selective editing is based on score functions. A record score is a combination of local scores for each of a number of important target parameters. Local scores are generally products of a risk component and an influence component. The risk component is measured by comparing a raw value with an "anticipated" value, often based on information from previous data. The influence component is measured as the (relative) contribution of the anticipated value to the estimated total. Only records with scores above a certain threshold are directed to interactive editing.

In *automatic editing*, data is edited by computers without any human intervention. We distinguish between correcting systematic errors and random errors, and different kinds of techniques are used to edit these errors. Once detected, systematic errors can often easily be corrected because the underlying error mechanism can usually be deduced. Random errors can be detected by outlier detection techniques, by deterministic checking rules that state which variables are considered erroneous when a record violates the edit rules in a certain way, or by solving an optimisation problem, for example by minimising the number of fields to change so that the adjusted data satisfies all edit rules.[7] With the introduction of automatic editing, one was able to clean relatively large amounts of survey data in a reasonable time.

Macro-editing can be used when (most of) the data set has been collected. It checks whether the data set as a whole is plausible. We distinguish between two forms: the aggregation method and the distribution method. The *aggregation method* consists of verifying whether figures to be published seem plausible by comparing them to related quantities from other sources. This method is often used as a final check before publication. In the *distribution method* the available data is used to characterise the distribution of variables. Then individual values are compared with this distribution.

deemed suitable for publication purposes. The tableplot for the corresponding edited data is shown in Figure 2.

The distributions of the numerical variables in Figure 2 are much smoother than in Figure 1; they are less disturbed by row bins with large values. In particular, the difference between the distributions for "results" stands out. The same is true for the categorical variables "sector" and "size". Both display a much smoother distribution in Figure 2, and in "size" the remarkable disturbance displayed in the upper part of the column in Figure 1 is completely gone. This is very likely the result of corrections for so-called "thousand errors": businesses have to report their amounts in thousands of euros, but many neglect to do so. Also, note that "book profit" no longer suffers from missing data and the negative "turnover" values are gone. These are all indications that editing has improved the quality of the data.[4]

A bigger problem

Having gained such experience editing large administrative data sets, we felt ready to process Big Data. However, we soon found out we were unprepared for the task. Owing to the lack of structure (variety) and the large amounts of data (volume), we discovered that several editing techniques developed for survey data cannot be applied efficiently to Big Data, including interactive editing and selective editing (see box for definitions).

Even automatic editing methods are hard to apply to Big Data as they often require subject-matter knowledge in the form of a detailed set of edit rules. Obtaining and applying such knowledge is challenging for many Big Data sources. The most promising traditional kind of automatic editing methods are those based on statistical modelling as these do not require user-specified edit rules. However, even these are hampered by the selectivity of many Big Data sources since not all parts of the target population may be equally well represented. This negatively affects the estimation of model parameters.

The aggregation method of the macro-editing approach, where the plausibility of publication figures is checked by comparing these figures to related quantities from other sources, can be applied to Big Data. The aggregation method is, however, only suited as a last final check before publication of the figures and should almost always be supplemented by other editing techniques that can be applied earlier in the cleaning process.

Visualisations developed for "merely" large data sets, such as the tableplot, do hold promise for Big Data and its three Vs. Volume can be dealt with by binning or aggregating the data. Velocity can be addressed by making animations or by developing a dashboard. Variety can be handled through interactive interfaces that allow visualisations to be adapted quickly. Besides the tableplot, other promising visualisations are "treemaps" and "heatmaps".[4,5] Such visualisations can often be used to monitor the effects of the editing process. However, to correct errors in Big Data sources, new approaches are needed.

Cleaning Big Data

The approach we describe here has been developed specifically for road sensor data. The sensors work as follows: whenever a vehicle passes by, information about traffic flows is generated, such as vehicle counts and mean speed of vehicles passing. In the Netherlands, for about 60 000 sensors, the number of passing cars in various vehicle length categories is available on a minute-by-minute basis.

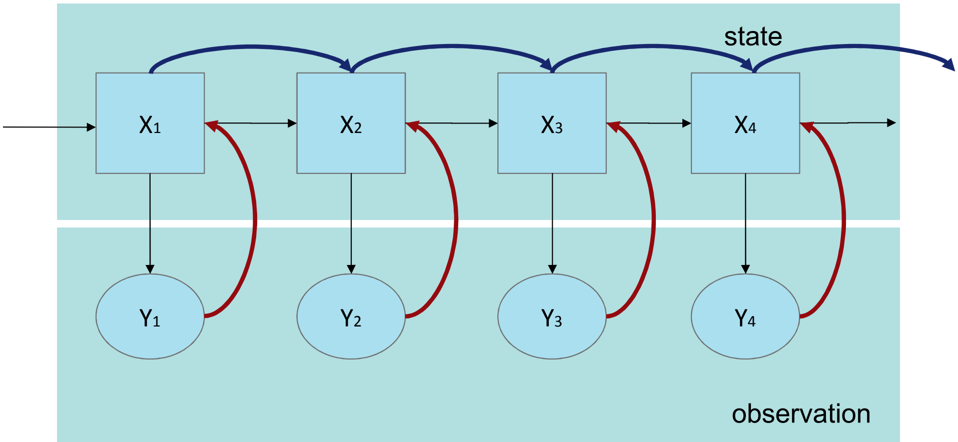


Figure 3. A Markov chain model for road sensor data

The most important issue we ran into while studying road sensor data was that the quality of the data fluctuates tremendously. For some sensors, data for many minutes is not available and, because of the stochastic, or random, nature of the arrival times of vehicles at a road sensor, it is hard to directly derive the number of vehicles missing during these minutes.[6]

The high frequency at which the data is generated severely hampers the use of traditional data editing techniques. Even traditional automatic editing and graphical macro-editing failed in this case. The breakthrough was the realisation that the high frequency of the data enables us to apply signal processing techniques for editing and imputation purposes. In particular, one can estimate a Markov chain model for each road sensor (see Figure 3).

In such a Markov chain model a road sensor can be in a certain state at time t, where a state is the number of vehicles that passed over the road sensor during the last minute. A Markov chain is a random process that undergoes a transition from one state at time t to another state at time $t + 1$ with a certain probability. The most characteristic aspect of a Markov chain is that it is memoryless: the probability of transitioning from the current state to the next depends only on the current state and not on the preceding states.

In Figure 3, $Y_t (t = 1, 2, \ldots)$ denotes the observed signal at time t, that is, the observed (but possibly incorrect) number of vehicles that passed the sensor during the last minute before time t, and X_t the true (unobserved) signal, that is, the true number of vehicles that actually passed the sensor. The observed data $Y_t (t = 1, 2, \ldots)$ is used to estimate the transition probabilities to go from one state X_t to the next X_{t+1}.

The most common kind of error that occurs in road sensor data is that observations are missing due to the fact that the sensor is temporarily not working properly. The Markov chain model can be used to automatically correct for this kind of error. Namely, in cases where the observed signal Y_t is missing, the Markov chain draws a value for X_t using the previous true state X_{t-1} and the estimated transition probabilities. The Markov chain model makes it possible to automatically edit and correct exceedingly large amounts of data. We applied this successfully to 105 billion records.

This might be a new era of Big Data, but the old requirements for robust and reliable data remain

Growing up

The use of Big Data for statistical purposes is still in its infancy, particularly in the development of efficient editing techniques. One of the big challenges for Big Data is monitoring the quality of the data without the need to inspect the data in its most granular form. As a result, one needs technological and methodological aids to inspect quality at an aggregated level.

An even bigger challenge is to detect and correct errors in Big Data quickly and automatically. The most promising direction appears to be the development of tailor-made automatic editing techniques such as the Markov chain approach we applied to road sensor data.

It is an exciting period for statistics, and official statistics in particular. Big Data offers the possibility of producing statistics in new ways by thinking "outside the box", and it will inevitably stimulate the development of new editing approaches. It might be a new era, but the old requirements for robust, clean and reliable data remain.

References

1. Lansdall-Welfare, T., Lampos, V. and Cristianini, N. (2012) Nowcasting the mood of the nation. *Significance*, **9**(4), 26–28.

2. Daas, P. J. H. and Puts, M. J. H. (2014) *Social Media Sentiment and Consumer Confidence*. Statistics Paper Series No. 5, European Central Bank, Frankfurt.

3. Daas, P. and Puts, M. (2014) Big Data as a source of statistical information. *The Survey Statistician*, **69**, 22–31.

4. Tennekes, M., de Jonge, E. and Daas, P. (2013) Visualizing and inspecting large datasets with tableplots. *Journal of Data Science*, **11**, 43–58.

5. Tennekes, M., de Jonge, E. and Daas, P. (2012) Innovative visual tools for data editing. Presented at the United Nations Economic Commission for Europe Work Session on Statistical Data Editing, Oslo. bit.ly/1GMhemJ

6. Daas P. J. H., Puts, M. J., Buelens, B. and van den Hurk, P. A. M. (2013) Big Data and Official Statistics. *Journal of Official Statistics* **31**(2), 1–15.

7. Fellegi, I. P. and Holt, D. (1976) A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, **71**, 17–35.

Marco Puts is statistical researcher at Statistics Netherlands, with a focus on Big Data processing and methodology. His special interests lie in the field of artificial intelligence and data filtering

Piet Daas is senior methodologist and project leader for Big Data research at Statistics Netherlands. His main fields of expertise are the statistical analysis and methodology of Big Data with specific attention to selectivity and quality

Ton de Waal is senior methodologist at Statistics Netherlands and professor in data-integration at Tilburg University. His main fields of expertise are statistical data editing, imputation and data-integration

Draft September 4 2015

Access to New Data Sources for Statistics:

Business Models for Private-Public Partnerships

Workshop jointly organised by PARIS21 and OECD

Date: 17-18 December 2015, OECD Headquarters

Background:

> *"For the 1.25 billion people who live in the rich countries of the world, [...] people are concerned not about generating more data but rather coping with a deluge of it. For many of the remaining more than 80% of the world's people [...], there is a different concern – being counted in the first place. Two thirds of birth are not officially registered in least developed countries, while causes of death remain largely un-known – despite both information are critical for better policy making" -- A Roadmap for a Country-Led Data Revolution, PARIS21*

> *"More and more organisations are starting to leverage large volumes of digital data generated from myriad transactions and production and communication processes. These large streams of data, that are now commonly referred to as 'big data', are generated through information and communication technologies, including the Internet as well as ubiquitous, wired sensors that are now capturing activities in the physical world. […] The analysis of big data, increasingly in real time, is driving knowledge and value creation across society" (OECD 2015 Data-Driven Innovation).*

Both quotes describe the challenges and opportunities of new innovative data sources that have gathered significant momentum in the past few years. Rising budgetary pressure on National Statistical Offices (NSOs) and other public data providers comes along with a desire to minimise the response burden for firms and households, declining response rates from some traditional survey sources and strong demand for real-time information as well as a requirement to produce statistics and information more efficiently and more timely.

It is increasingly recognised[1] that traditional statistical approaches may have to be complemented by exploiting new, innovative data sources to "[…] increase significantly the availability of high-quality, timely and reliable data disaggregated by income, gender, age, race, ethnicity, migratory status, disability, geographic location and other characteristics relevant in national contexts[2]".

Some of the new data sources appertain to the private sector, in particular telecom data, data from social media and credit card records. There are already many instances where official statistics use

---

[1] See in particular the Reports *A World That Counts* by the United Nations http://www.undatarevolution.org/report/ and *Informing a Data Revolution* by PARIS21 http://datarevolution.paris21.org/ .

[2] Final draft of the outcome document for the UN Summit to adopt the Post-2015 Development Agenda, July 2015: "Transforming Our World: The 2030 Agenda for Sustainable Development".

commercial, non-survey data sources for the construction of official statistics (for instance scanner data in price statistics, credit card information for balance of payments) and the potential is growing. A recurring issue is if and **how such data can be accessed by NSOs for statistical purposes** as there are **legal, institutional and commercial aspects** to using the relevant information. Issues of **governance** arise as well as questions about the **business case for** those **private actors** holding the data.

The OECD and the Partnership in Statistics for Development in the 21st Century (PARIS21) propose to organise a joint workshop on this topic.

The OECD has been active in exploring the possibilities of big data for some time, by bringing together analysis, evidence and experience from its members and beyond. Internally, the OECD has launched a Big Data initiative to explore how new data sources can be used for its own statistical and analytical purposes. The OECD, through its multi-faceted work, and its close links to the community of official statisticians, to governments at large as well as the private sector is well placed to examine the opportunities and challenges of using new data sources for statistics from various angles.

PARIS21 is hosted at the OECD and brings together national, regional and international statisticians, policy makers, development professionals and other producers and users of statistics. PARIS21 has a proven and strong track record in improving statistical capacity in developing countries. Its *Informing a Data Revolution* project focuses on how to improve data systems, which are crucial to generating the data needed to reduce poverty and to monitor the sustainable development goals. Access to and governance of new sources for official statistics is an issue that concerns developed and developing countries alike.

<u>Objectives and outcomes of the workshop:</u>

The workshop will start by reviewing some of the existing experience accessing and using new data sources (telecom data, social media, sensors, credit card records etc …). One particular objective is to address the question how to develop **business models** that are attractive for the private-for-profit and non-for-profit sector to provide access and share its data.

More broadly, the workshop aims at **bringing together practical experience regarding access to commercial data for statistical purposes**, and **identifying key legal, institutional and statistical issues associated with such access**. The discussions in the workshop would be summarised in a report and constitute part of PARIS21's and OECD's contribution to the work on the Sustainable Development Goals, directly feed into the OECD's internal Big Data Initiative, contribute to the work streams of the Global Partnership on Sustainable Development Data[3], the UN Global Working Group on Big Data for Official Statistics[4] as well as other international initiatives such as *ModernStats*[5].

[3] Global Partnership on Sustainable Development Data is a coalition of key stakeholders engaging to take action to harness the data revolution. The partnership will be officially launched at the UN General Assembly on see: http://www.one.org/international/press/harnessing-data-revolution-to-drive-sustainable-development/

[4]See http://unstats.un.org/unsd/statcom/doc15/2015-4-BigData-E.pdf

Format:

1 1/2-day expert workshop with about 50 participants with the following structure:

Session 1: New data sources for statistics: examples of current projects

Presentations from both developed and developing countries showing how new data sources have been accessed and used to complement official statistics. Presentations should in particular relate to how agreements were set up with private data owners.

Session 2: Statistical Office Perspective

Presentations to review statistical requirements in Public-Private Partnerships such as control over data quality, or the durability of data sources

The session would also highlight expertise that NSOs can offer in regards to ensuring confidentiality when micro data from NSOs' own sources or administrative sources are used.

Session 3: Business Models and Incentives

Presentations by the private sector and by academics to provide their perspective on incentives, obstacles and business models needed to make data access possible

Presentations by regulatory or international institutions (such as Eurostat) on possible governance models for individual Public Private Partnerships, and the need for legislation and regulation (or absence thereof)

Session 4: Wrap up and way forward

Participants:

PARIS21, OECD (STD, STI, ITN, ITF); NSOs, private sector, relevant international organisations (UNECE, Eurostat, UN big data group, ITU); academia

[5] *Modern Stats* is a group of NSOs, set up under the auspices of the UN Conference of European Statisticians that co-operates on the modernisation of statistical production and services; see: Www1.unece.org/stat/platform/display/hlgbas.

Uses of Big Data for Official Statistics: Privacy, Incentives, Statistical Challenges, and Other Issues

Discussion Paper

by Steve Landefeld
Senior Advisor to the United Nations Statistics Division

International Conference on Big Data for Official Statistics
Beijing, China, 28 – 30 Oct 2014

United Nations Global Working Group on Big Data for Official Statistics
Beijing, China, 31 Oct 2014

Abstract

This paper provides an overview of big data and their use in producing official statistics. Although advances in information technology, data sources, and methods have driven interest in the use of "big" sets of business and administrative government data collected and used for non-statistical purposes, use of such data is not new. Nor is it likely to be a panacea for statistical agencies confronting demands for more, better, and faster data with fewer resources. However, with careful attention to incentives, protection of privacy, and integration of these non-statistical data with existing statistical data, big data can play a large role in improving the accuracy, timeliness, and relevance of economic statistics at a lower cost than expanding existing data collections.

Table of Contents

Uses of Big Data for Official Statistics: Privacy, Incentives, Statistical Challenges, and Other Issues

A. The Growth and Potential Impact of Big Data

1. Uses of "Big Data" were recently heralded in a U.S. White House report as "fundamentally reshaping how Americans and people around the World live, work, and communicate."[1] Examples include saving lives through epidemiological research using big data from neonatal intensive care units; tracking the incidence of flu through geographic analysis of Google searches on the use of the word "flu"; making the economy work better through the analysis of delivery truck GPS data to develop more efficient delivery routes; and saving taxpayer dollars by identifying patterns of fraud in medical care claims.

2. Big data have also been described as a transformative tool for official statistics,
The statistical community has recognized the potential for big data in improving accuracy and reducing costs for official statistics. In 2014 the United Nation's established a global working group to:

> "provide a strategic vision, direction, and a global programme on big data for official statistics, to promote practical use of sources of Big data for official statistics, while finding solutions to their challenges, and to promote capacity building and sharing of experiences in this respect."

3. Examples of the use of business and administrative for statistical purposes include the "scraping" of internet data to produce the "billion prices" Consumer Price Index; the use of payroll data from the Automatic Data Processing Company (ADP) for its monthly employment index; the use of international postal data to create an International Letter-Post Index that can be used as a leading index or to improve forecast accuracy, and the use of Google searches for "now-casting" of the state of the economy.

4. Several factors have facilitated these advances in the use of big data. Among the more important of the factors are:

- Advances in information technology that have lowered data collection, storage, and processing costs.
- The development of new sources of data and improved access to existing big data sets, on and off-line,

[1] Big data is defined in this paper as the use of large-scale business and administrative data sets that are being used for secondary purposes other than those for which the data was originally collected.

- The parallel development of creative and powerful new methods to exploit "big data."

- The recognition --- through some of these high profile projects -- that we have massive stores of data collected for such purposes as business, administration, health care, meteorology, and traffic that can be used, alone or in combination with other data, for an array of purposes other than those for which they were originally collected.

B. Uses of Big Data for Official Statistics

5. Although a large share of official statistics are based on official surveys, there is a long history of the use of non-survey data in the area of National Accounts, which have been described as a mosaic of public and private data, with most of it originally collected for purposes other than their use in constructing national accounts statistics.

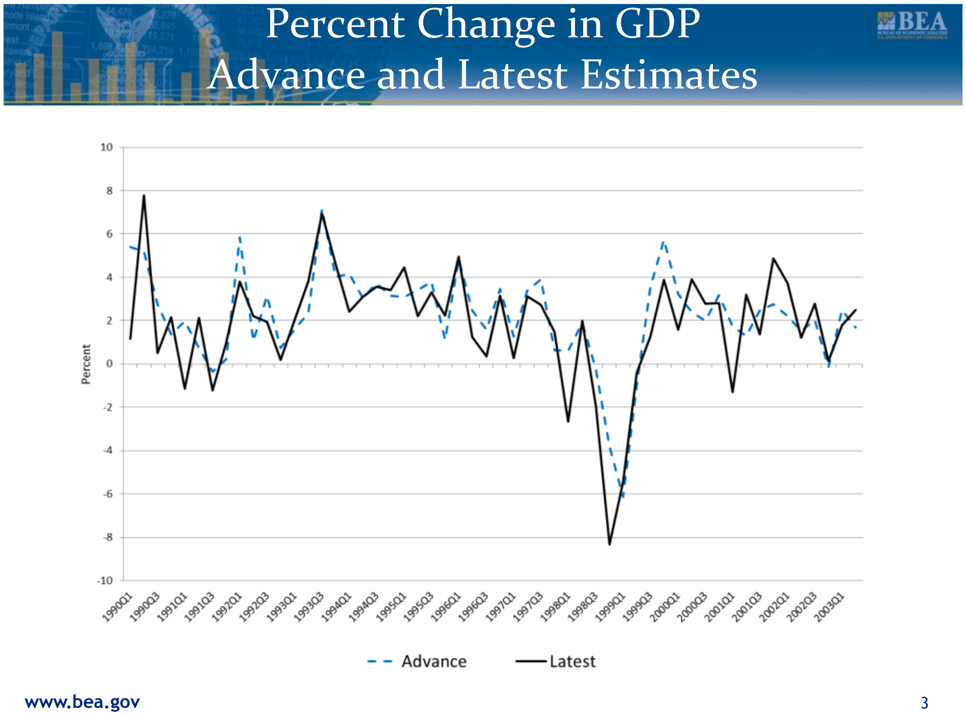
6. Indeed, since their inception, national accounts have used a mix of public and private data to provide a comprehensive picture of overall economic activity that is timely and accurate. The United States and other countries make extensive use of partial data – public and private – as extrapolators for its early estimates. Most of the non-statistical data used have been aggregations of business and government micro-data, although micro-data is used in matching of statistical and non-statistical data to improve official statistics by developing bias adjustments for survey data, improving coverage, and by identifying reporting and other problems.

7. For privacy and other reasons this pattern of the use of big data collected for non-statistical purposes -- as extrapolators and methodological research and improvement tools -- is likely to continue.

8. Because these business and administrative data are collected for non-statistical purposes, they usually do not meet statistical standards in terms of representativeness, concepts, definition, collection methods, etc. To use these administrative and business data national accountants must investigate and understand the statistical characteristics of the data and improve the accuracy of these non-statistical extrapolators through weighting, filling in gaps in coverage, bias adjustments, averaging with other extrapolators, and benchmarking and balancing.

9. For most periods, these extrapolators have worked well and at the same time lowered costs relative to a system of ongoing surveys that collected data designed just for national accounts purposes. As can be seen from Chart 1, the early estimates using "mixed data" provide a timely and accurate general picture of economic activity. The early GDP estimates, based on a mix of public and private extrapolators, released roughly 30 days after the end of each quarter track the later estimates, based on benchmark official data, well.

Chart 1



10. Examples of the source data used for the early GDP estimates are official statistics such as U.S. Census Bureau monthly retail sales, shipments, and inventories data and BLS employment data. All are based on early sample results that will subsequently be revised. Where official monthly indicators are not available, other government and private sources are used. Examples of the private source data aggregations included in the accounts are:

- Ward's/JD Powers/Polk (auto sales/price/registrations)
- American Petroleum Institute (oil drilling)
- Airlines for America (airline traffic)
- Variety magazine (motion picture admissions)
- STR (hotels and motels)
- Investment Company Institute (mutual fund sales)

11. In evaluating business and administrative data for use in national accounts, one of the first questions to be asked is how closely do the data fit with national accounts concepts? A leading example of the impact of differences in concepts is found in the differences in profits using the accounting rules for business profits, tax profits, and economic profits. One would expect that use of the profits from what must be two of the largest "big economic data" bases (U.S. corporate reports and tax filings to Securities and Exchange Commission and the Internal Revenue Service) would make corporate profits one of the most reliable and accurate components of the national accounts. As it turns out, they are one of the most volatile, most revised components of national accounts.

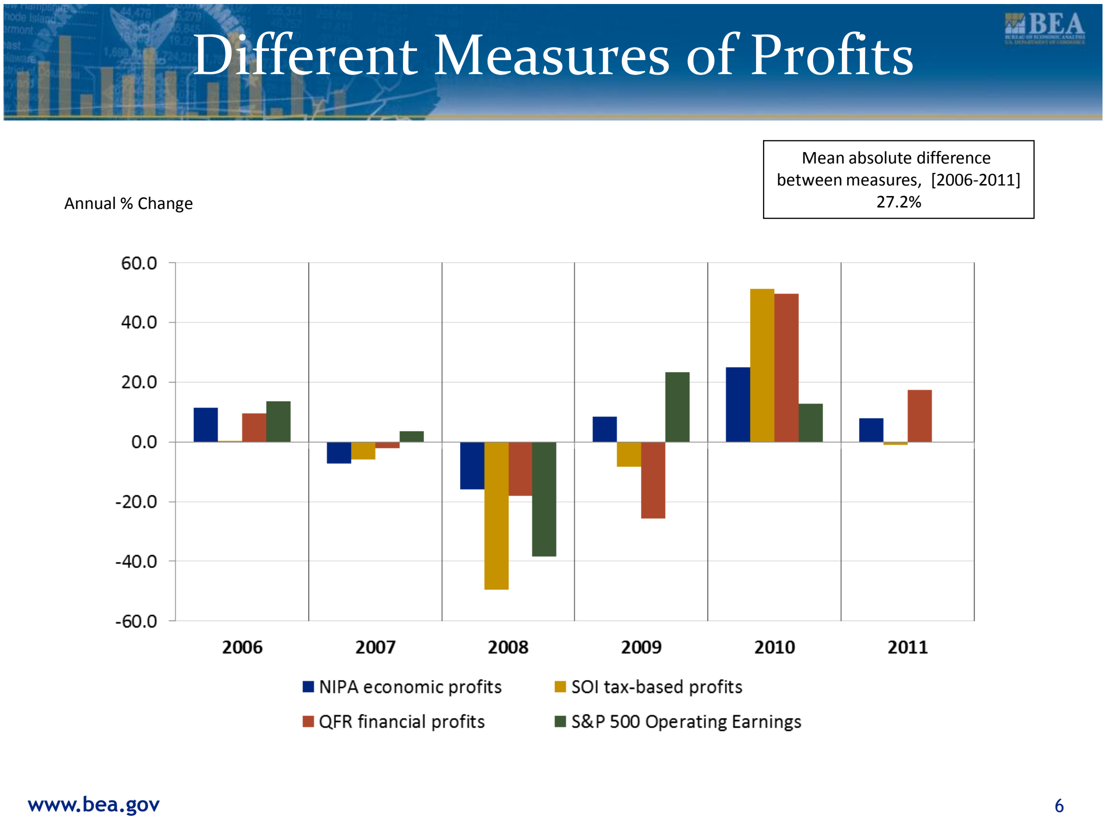
12. Business profits are based on rules such as those promulgated by the U.S. Financial Accounting Standards Board (FASB) and those laid out in the Internationals Financial Accounting Rules (FARS). Profits reported by those same firms to tax authorities use tax accounting rules that include incentives for investment, such as accelerated depreciation, or investment tax credits. Economic profits adjust for inflation and accelerated depreciation to value inventories and depreciation at their "true" replacement cost, and deduct capital gains and losses so that profits reflect the profits earned from production in the current period. Also, the coverage in each of the data sets differs.

13. As can be seen from Chart 2, the differences between growth rates in the three measures can be very large. Although for must years the different measures produce similar changes, periods where there are differences, the differences are quite large. In 2009 the different measures produced estimates that ranged from plus 20 percent to minus 20 percent. Further, tax returns and profit reports can be revised after the initial filing to reflect carry-forward and carry-back provisions for such items as operating losses, research and experimentation credits, as well as revisions based on IRS reviews and audits of their initial filings for up to 10 years after the initial report. As a result of these differences, estimates based on these data can be hard to interpret and are subject, even in the aggregate, to large revisions. Revisions to specific industries can be especially large and have a significant impact on industry profits, output, and supply and use estimates.

14. A second question that must be addressed in the use of these big data is the consistency of the time frame in the source data with the time frame for the national accounts estimate. In another example from the United States, data from a large national monthly payroll survey by the Bureau of Labor Statistics, are used by the Bureau of Economic Analysis to estimate monthly and quarterly compensation. One of the adjustments to the payroll data is for timing, which can be especially important when a major strike occurs during the week covered by the monthly payroll survey (or during a week(s) not covered by the survey). Other examples, include difficult detailed micro-data adjustments to state and local and business data from a fiscal to a calendar-year basis.

15. The third issue relates to the representativeness of the external source data and any selection biases that may be present in the data. For example, in the case of financial reports, they are limited to publicly-held corporations, and exclude important privately-held companies as well as the business income of partnerships and sole proprietors. This is a significant problem because the behavior of the included vs. excluded data can differ markedly over the business cycle and by type of industry (retail or services vs. manufacturing).

Chart 2



16. Unfortunately, some of the largest gaps in coverage for the official economic statistics are difficult to fill using publicly available data. Significant gaps in official output statistics such as those in services and local governments are hard to fill because they are in sectors dominated by a large number of relatively small units using an assortment of concepts, definitions, reporting periods, and accounting rules. Small firms do not file public reports or make available anything other than their industry, services offered, and location (items normally found in business directories), or sales advertised on the internet. Because they are the among the items businesses find most sensitive, small firms generally do not post or file information on their sales, prices, and costs. Filling gaps in local government data is also difficult. While counties, townships, and other small governments provide taxpayers and voters with financial reports few are on a consistent or comparable basis (e.g. reporting period, accounting conventions, etc). Similarly, gaps in income statistics are in hard-to-fill areas like small business income, which is often only reported on individual tax returns.

C. Process for evaluating and using new big data:

17. Incorporating new extrapolators is a multistep process. The first step is evaluating the concepts, definitions, coverage, and performance of extrapolators relative to more comprehensive

and consistent and annual and benchmark data. The second step is developing new methods to use new extrapolators, including benchmarking, weighting and combining with other indicators; bias adjustments. These new extrapolators then need to be evaluated relative to existing extrapolators and benchmarks to access their accuracy. The final step is developing seasonal adjustment factors for the new extrapolated data, which is a difficult process given the normally short time series for the new data.

18. One of the major challenges of this process, is the need to avoid, wherever possible, the use of complex econometric techniques. In general, econometric forecasts and extrapolation add little to accuracy and often do not perform as well in large scale forecasting or as extrapolators as simple trend extrapolations. Complex econometric models also are difficult to understand and assess and result in a loss of transparency to users of the data. Further, if the models are complex or produce only aggregated results, there will be a loss of drill down capacity and links to key indicators.

D. Loss of Control in Using Big Data

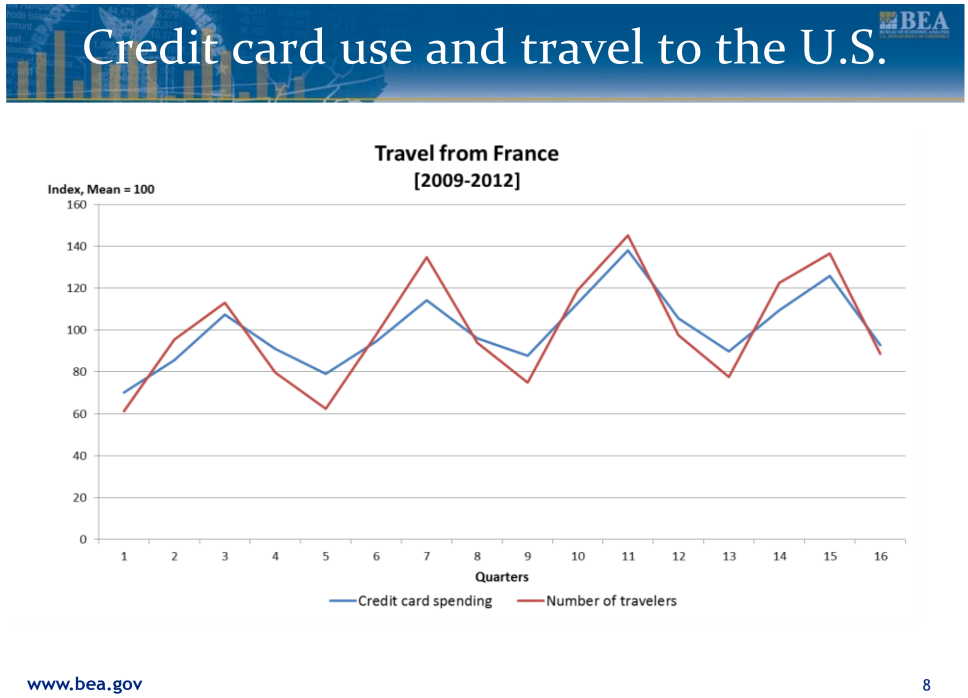
19. In addition to the challenges cited above, the use of big data results in a certain loss of control and dependency on the part of official statisticians. If a company or government agency decides that for business reasons to change definitions, collect different data, or entirely stop their data collections, official statisticians have virtually no leverage to prevent the loss of such data. Although, with increasingly tight budgets, official surveys and statistical sources are subject to discontinuation, official statistical agencies at least have some degree of leverage and control and the ability to reallocate funds to the highest priority statistics.

20. In the United States the challenges associated with depending on business and administrative data has been manageable, but requires having benchmarks and official statistics as the baseline for extrapolations using unofficial data and on the flexibility to change source data and methods relatively quickly.

E. Examples from some preliminary collaborative research and analysis

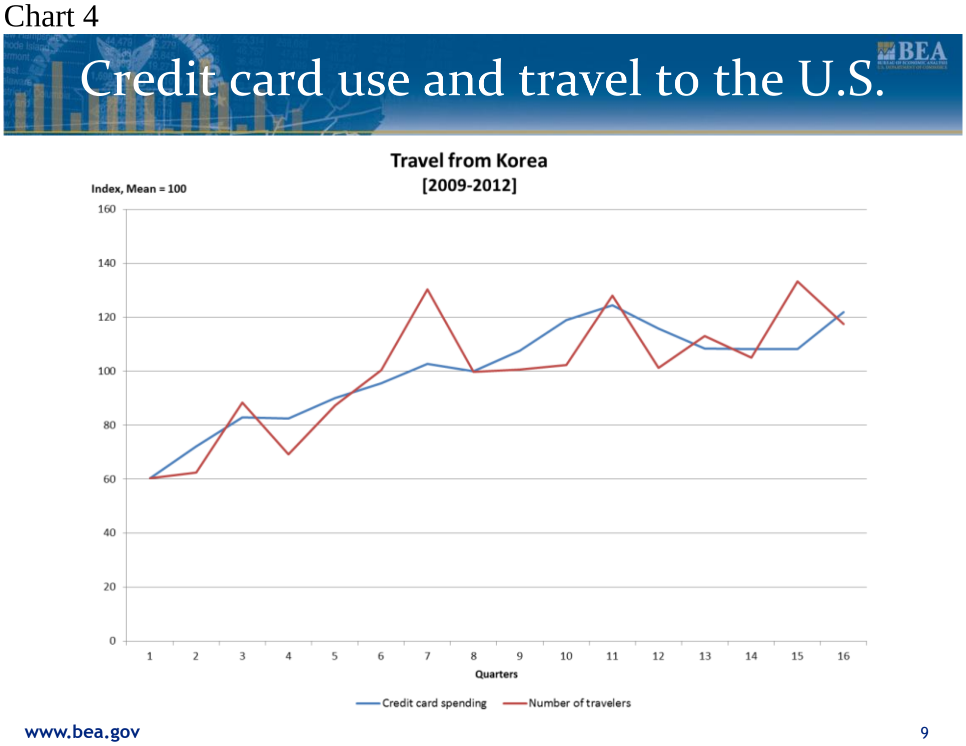
21. The following charts illustrate the challenges in using big data. There are often huge numbers of observations, but they may not be representative observations. Chart 3 compares data on credit card use by U.S. citizens in France with customs data on the number of U.S. citizens travelling to France. Neither is the appropriate measure that one would want to measure spending by all U.S. citizens travelling to France. The credit card measure covers credit card spending by those using credit cards in France and the number of travelers covers all U.S. citizens traveling in France. Although, the two series produce a similar pattern, the credit card pattern produces a more striking seasonal pattern that may be hard to explain. Does spending per traveler have a seasonal pattern, or is the spending pattern by credit card different than the pattern of cash spending?

Chart 3



22. Chart 4 shows credit card data and traveler data for travel from South Korea to the United States and the patterns are quite different. The data may be useful as a long-term trend indicator, but without weighting of credit and cash uses, and development of indicators for cash use, it will be hard to use the credit data for measuring monthly and quarterly patterns.

Chart 4



23. Chart 5 and 6 show data from a popular household budget tracking "app" that covers credit card and all other spending. The data from the tracking service fits quite closely with the representative data from the official U.S. Census Bureau data on retail trade for clothing. However, the same comparison for electronic goods shows that the budget tracking data significantly understate seasonal peaks in spending. Perhaps households that sign up for a budget tracking service are less prone to holiday "binge" shopping for "big-ticket items."

24. Chart 7 shows a measure of small business activity based on a U.S. small business accounting software, labeled alternative net profit indicator in the Chart. The other two indicators labelled NFBI (Non-financial business indicator) are measures used in the U.S. national accounts based on official employment and tax data, as well as other indicators for key small business sectors. (One of the two NFBI excludes capital and inventory gains and losses.) Small business income is extremely hard to track, even with official statistics, and any new data can be a big help, but in this case the trends from the accounting software data for the most recent years of the economic recovery, are so sharply at odds with the official data that it is hard to figure out how to use them. Although, they measure different things it is difficult to square the improving official receipts and employment numbers for small business with the falling net profits from small business coming from the accounting software data.

Chart 5

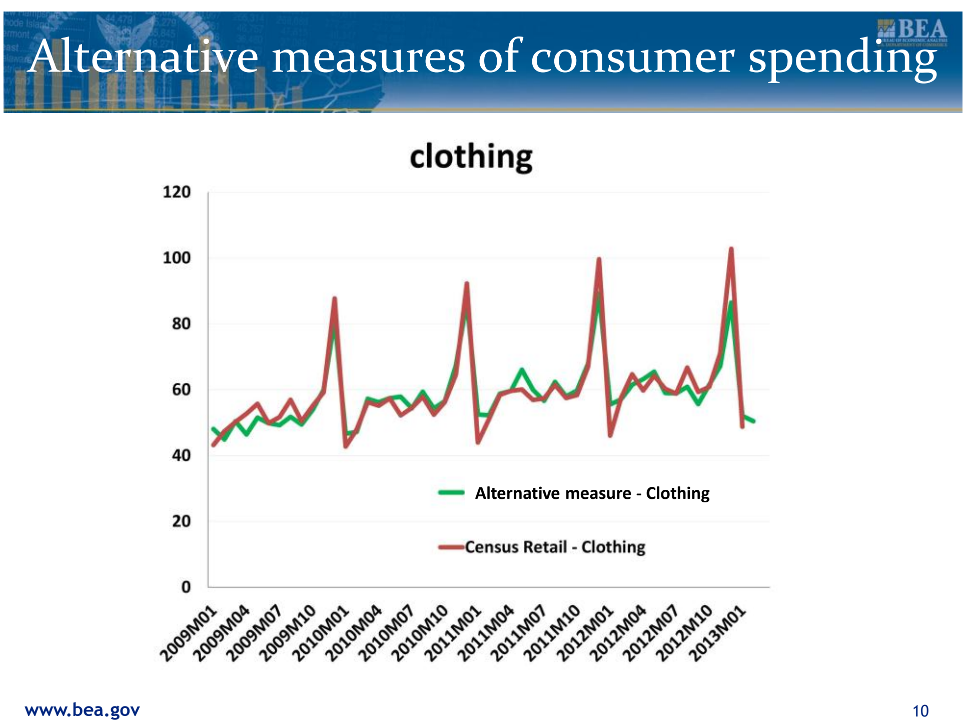


Chart 6

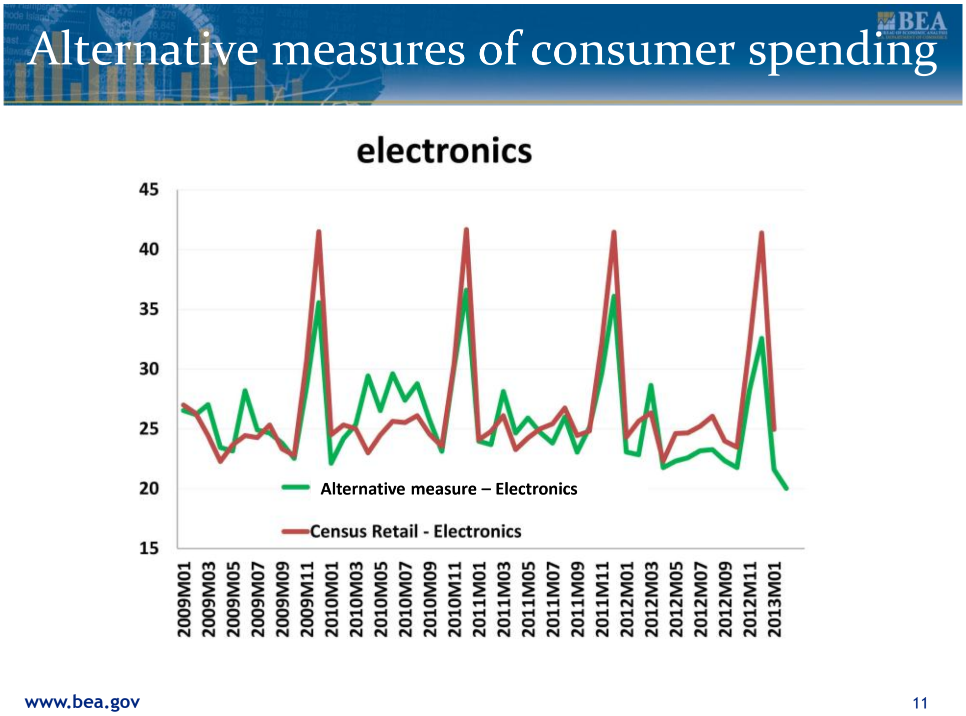
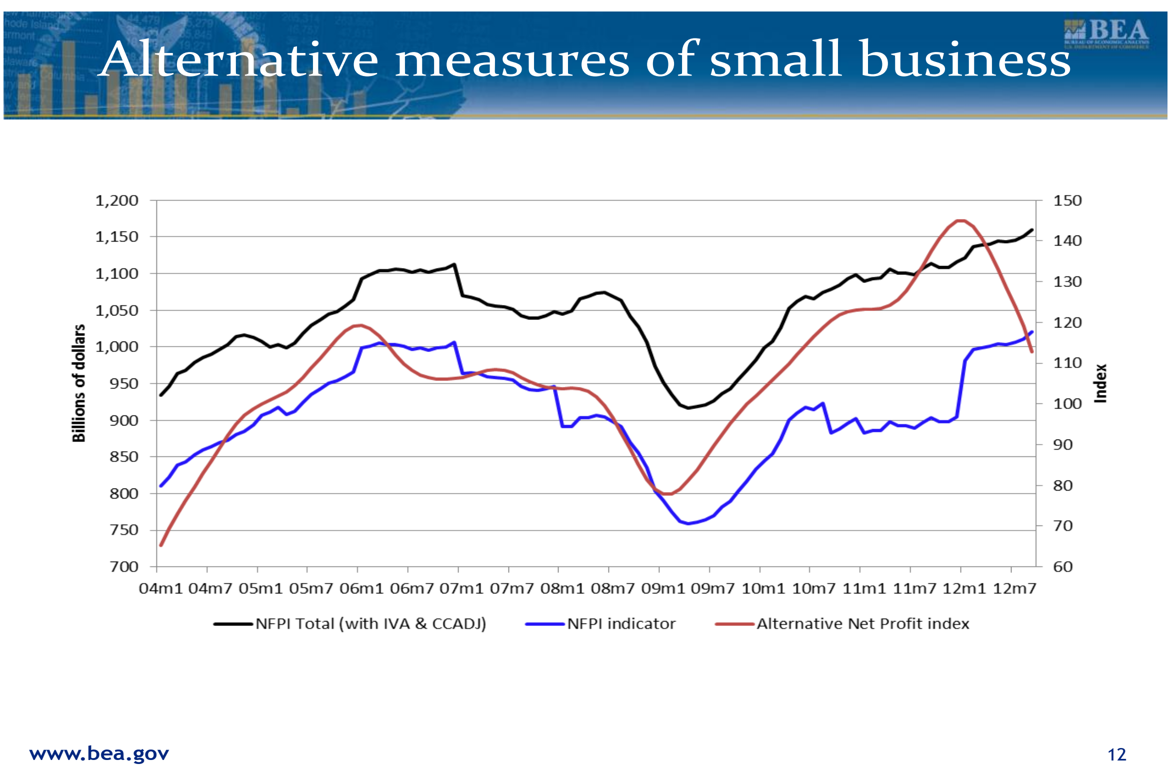
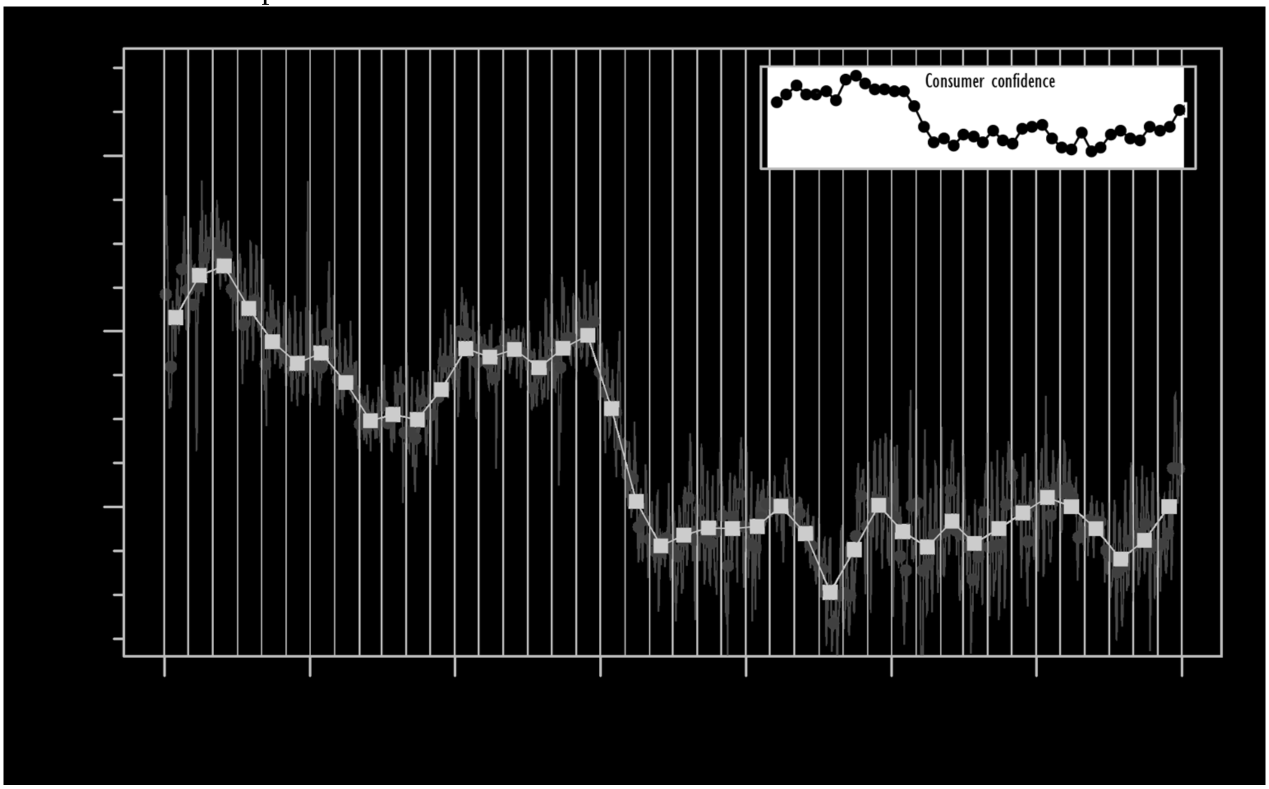


Chart 7



25. Chart 8 shows the relationship between Dutch social media sentiments on the state of the economy (Facebook supplemented by twitter messages) and monthly consumer confidence. The high correlation over the three-and- a-half years studied (2010-2013) provide a good example of how such data could supplement, and or be integrated with existing indexes.

Chart 8: Relationship Between Dutch Consumer Confidence and Social Media Data



Source: Piet J.H. Daas and Marco J.H. Puts, "Social Media and Consumer Confidence, " European Central Bank, Statistics Paper Series, No. 5 September 2014

26. Other examples include an Indonesian study of Facebook and twitter data on expressions regarding food prices and food price inflation and the potential for the use of social media for official statistics including the need for validation, methods for filtering out noise in the data gathered from social media, and developing robust estimates for their use with official statistics.[2]

27. Another study by Eurostat to assess the feasibility of the use of mobile positioning data for measuring tourism cited their large potential but pointed to the need to resolve a number of issues including: privacy and regulatory issues related to privacy; public opinion relating to the use of mobile data; financial and business related barriers to access; technical issues regarding use of and access to mobile data; and methodological issues relating to effectively using mobile data.[3]

[2] Pulse Lab Jakarta, "Mining Indonesian Tweets to Understand Food Price Crises," UN Global Pulse, Methods Paper, February 2014.

[3] EUROSTAT, "Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics, Consolidated Report Eurostat Contract No 30501.2012.001-2012.452, 30 June, 2014.

28. The final example is the use of postal big data. Despite declining use of the postal service around the world, Anson and Heble show how postal data can be used to develop an International Letter-Post Weight Index that seems to work well as a leading indicator catches changes in direction before a commercial leading index over the two year period 2010-2012 as well as differences in performance across countries in Europe. They also suggest that such real-time data could improve forecast accuracy. As with the other studies, use in leading indicators or forecasts would require testing over a longer time frame, developing methods that would explore not only direction of change, but relative magnitude of changes, and testing across countries.

F. Incentives: Exploiting Public and Private Benefits of New Extrapolators

29. The potential benefits to official statistical agencies and owners of public and private administrative data are large. For official statistical agencies, if the challenges of concepts, definitions, and representativeness can be resolved, the use of big data has the promise of more timely and detailed data at a significantly lower cost than new or expanded survey collections.

30. For example, in the United Sates, mounting a new survey collection may cost well over $20 million, whereas a micro-data matching exercise using existing data that produces more timely and detailed BEA may cost less than 1/5 of the cost of a new survey. Indeed, a project that matched existing firm level data with existing plant level data expanded the industry detail on U.S. foreign direct investment data from roughly 100 industries to over 600 industries and from national level data to data for all 50 U.S. states at a cost of $3 million.

31. EUROSTAT has matched existing business register (public and private) and trade data at the micro and aggregate level to create detailed data on the characteristics of importing and exporting firms that are extremely helpful to designing trade and investment policies.

32. The owners of public and private data want to exploit the data they collected for business and administrative purposes for other purposes. For businesses, these other purposes include marketing, geographic location plans, short and long-term investment plans, and the benchmarking of company to industry performance. For the public sector, these other purposes in include geographic planning for infrastructure and the provision of services; for understanding behavioral responses and characteristics of the population for designing health and unemployment insurance, for tax and other policies, for tracking the incidence of disease for public health purposes; and for evaluating the effectiveness of government programs.

33. For big data to be useful for all these secondary purposes the administrative and business data must meet the measurement purpose for which it is to be used. For example, is the number of Facebook uses of the word flu a reliable indicator for the incidence of flu? Or are the health records of the members of a large health insurance plan in an urban area representative of the general population, insured and uninsured? Or in the area of economic statistics, are on-line price

quotes representative enough of all prices to predict the official CPI with enough accuracy to be useful to traders in financial markets?

34. Official statisticians and big data users and producers, therefore, all benefit from harmonization of the various sets of statistics for secondary uses and have strong incentives to collaborate in doing so. As shown from the experience of national accountants, after examining the characteristics -- definition, population, methods, and performance -- of the source data they can often be benchmarked, bias-adjusted, and weighted (for use with other data) to successfully project official statistics.

35. There are also benefits from collaboration between official statistical agencies and information firms such as Google, Baidu, and Yahoo. Search engines may have problems identifying the most timely and authoritative data and may send users to dated secondary sources. By collaborating with official statistical agencies they can provide their customers with more accurate, timely, and relevant data searches. By linking to the original source data that customers are seeking they help their customers by raising their "statistical literacy" in understanding the implications of matches of what appear to be similar data from searches and "mash-ups." They also help official statistical agencies by providing improved access to, and understanding of, official statistics through broader internet dissemination and "branding."

36. Some of the most successful examples of the use "big data" have come from joint projects between the owners of private data and official statistical agencies. Such joint projects provide significant mutual benefits not realized by unilateral use of private source data. Examples include collaboration between the U.S. Bureau of Economic Analysis and IBM in the development of hedonic computer prices, with Chrysler and other motor vehicle manufacturers in developing new auto and truck pricing data, and with Google in the development of regional search engines.

G. Successful public-private collaboration requires

- A recognition of the mutual benefits of such a collaboration
- Transparency between the official statistical agency and owner of the business or administrative data regarding data collection and estimation methods used to produce the data; and
- Clear and strong rules for protecting the confidentiality of the data and of the proprietary methods used to produce the business or administrative data

37. Although seemingly obvious, the infrequency of public-private collaboration in some countries may inhibit such collaboration and require that statistical agencies initiate and outline the benefits of such collaboration.

38. Further, although transparency and understanding the data is critical to effective benchmarking of the private data to the fully representative official statistics, providing transparency on source data and methods may be difficult for proprietary data suppliers. Protection of the source data and methods are essential to the firm's ability to sell the data product to customers. Without protection of their intellectual property -- which is usually done through trade secrets rather than patents or copyrights -- competitors may be able to replicate the private data product without the development costs borne by the original owner, and undercut their sales.

39. Finally, protection of the private data, much of which is confidential customer data, is essential. Unfortunately, long-standing lack of trust in government in the area of privacy is a problem to such collaboration that has been exacerbated by recent highly public examples of governments' accessing private "big data" for National Security reasons (with or without the private owner's consent). While none of the recent transgressions, or indeed most past data breaches, have involved official statistical agencies, businesses and the public are inclined to mistrust government in general. Further, cases where official statistical agencies have made inappropriate use of confidential data have been quite public and likely tarnished the general reputation of statistical agencies around the world.

40. In addition to the use of collaborative projects to access confidential public and private data, patent and copyright protection may require their use even in the use what appear to be publicly available private data. Even if legal protections do not bar the use of public data through "data scrapping," collaboration that helps in understanding the characteristics and proprietary methods used to producing the internet and other data may be critical to the successful use of that data by official statisticians.

H. Privacy Concerns about New Uses for Big Data

41. As noted above, advances in the use of "big data", including high-profile political and national security access to big data, have raised significant privacy concerns. For individuals the concerns relate to the disclosure of detailed personal medical, financial, legal and other sensitive data; uses that would lead to discriminatory outcomes; and uses for tax, investigatory, legal, and other governmental purposes.

42. For businesses, the concerns are disclosure and release to the public of: commercially valuable marketing and other data sets; proprietary information on the methods and sources used to produce those data; disclosure to competitors of important strategic information on pricing, costs, profits, and markets; and the use of such information for tax, regulatory, investigatory, legal, and other purposes.

43. These concerns are not new concerns, nor is the use of data collected for business and governments for non-statistical purposes in the production of official statistics. In addition, many of the same confidentiality and privacy concerns have arisen in the course of centuries of government surveys of households and businesses. The mechanisms for addressing those concerns may be able to be carried over to rules, or protocols, for protecting data in today's big data world.

44. In developing such rules, protecting privacy and confidentiality are key. For business data - the following types of information must be protected:

- The data itself as an information product (micro and macro). This intellectual property of firm has economic value and can be sold. Government must make sure there is no disclosure that would give it away free.
- Data on details of businesses, such as prices, costs, and market share, that would be useful to competitors.
- Personal information on customers. Loss of such data through security breaches or hacking undermines the reputation of the firm, and discourages use of electronic transactions that can result in the loss of business.
- Proprietary information on the methods and sources used to produce the data

45. Such protocols to protect privacy are essential, and have been used for years-- to promote trust and address concerns that government may use of micro-data for regulatory, tax, and other policies. An erosion of public trust can reduce response rates on official surveys, reduce honesty in reporting, and reduce the overall accuracy of the official statistics collected from the business community.

46. The public also need to have their concerns addressed and their data protected. They need to be sure that there will be no disclosure of:

- Name and address or other identifying information that could be used for marketing and other business purposes.
- Intimate personal details, including such information as marital and health status, and income.
- Any information whose use could lead to discriminatory outcomes in such areas as employment, eligibility for loans, or eligibility for government programs.
- Any disclosure to non-statistical agencies that alters the balance of power of power between individuals and government, including the use of data collected for statistical purposes for tax, regulatory, investigatory and other non-statistical purposes.

I. Data Protocols for Public-Public and Public-Private Collaboration

47. Uses of Public Administrative and Business Data include both publicly available and confidential data. Data protocols focus on confidential data, although, as noted above they may be useful in fostering collaboration, understanding, and effective use of publicly-available administrative and business data.

48. A protocol for data access should begin by describing the benefits of such a collaboration. For example, a recent U.S. Executive order providing guidance in the use of protocols that would promote the use of Administrative data to leverage and improve statistical data, notes that, "...high-quality and reliable statistics provide the foundation for the research, evaluation, and analysis that help the Federal Government understand how public needs are changing, how well Federal policy and programs are addressing those needs, and where greater progress can be achieved."[4]

49. The U.S. memo then goes on to emphasize the importance of developing strong data stewardship policies and practices for the use of administrative data collected for non-statistical purposes (protection of privacy and confidentiality); documenting the key statistical attributes and quality controls for use of data; and developing memorandums of understanding (contracts or protocols) that cite relevant laws, regulations, policies, practices, responsible parties, and penalties associated with use and misuses of the data.

50. A protocol for access to private data should first cite the purpose of the agreement. It should cite the specific benefits (see incentives cited above) accruing to the public and private partners and the specific data products that will be produced by the agreement.

51. Second, the protocol should address the uses of the data and the quality of the data. Such an understanding helps prevent difficulties later in the project. It also provides a better understanding of the baseline data's strengths and weaknesses that will need to be addressed by benchmarking, weighting, and for seasonal, bias, and other adjustments.

52. Third, the protocol should cover the roles and responsibilities for the protection of the data. Elements that should be covered include confidentiality and privacy, data security; data transfer, media, and methods for transmission of data. It should also set out the specific penalties for unauthorized disclosure of information, including any applicable privacy laws and their penalties, up to, and including imprisonment.

[4] MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES, FROM Sylvia M. Burwell, Director of the U.S. Office of Management and Budget, "Guidance for Providing and Using Administrative Data for Statistical Purposes," February 14, 2014.

53. Fourth, the Protocol should cover such key details as the parties to the agreement; duration of the agreement; legal and programmatic authorities; estimated costs and payment (or cost sharing agreement, including in-kind services); administrative and programmatic contacts; procedures for the resolution of conflicts; procedures for reviewing, updating, modifying, cancelling, and renewing the agreement.

54. A copy of the elements included in a model agreement developed by the U.S. Office of Management and Budget that is the basis for the above recommendations is included below.

J. Other Roles for Official statisticians in the Use of Big Data

55. Some have suggested that official statisticians might play a role in auditing and certifying the accuracy of "big data" for official and private uses similar to the role played by consumer testing and ratings services such as Consumer Reports. Such a proposal was put forth in the United States and encountered stiff resistance from data users and the business community.

56. There were several reasons for this resistance. First, information products are an important source of revenue for firms and they do not wish to risk disclosure of the proprietary methods and source data they use to produce the data that they sell (see customers' privacy discussion above). Second, at the time and more so now, firms and their customers simply don't fully trust governments and are concerned that the data will be used for non-statistical uses (tax, regulatory, or investigatory); Third, there is a general resistance by industry to expanded government oversight and a strong preference for voluntary oversight. Fourth, such a role for official statistical agencies may well produce an adversarial environment with the very businesses that official statisticians must rely on for their regular survey data. Such an environment could well end up weakening, rather than strengthening official statistics by lowering response rates, and reducing the accuracy of responses.

K. Don't under-estimate the value of existing statistical and administrative data!

57. Despite the high level of excitement surrounding the use of on-line search data, internet prices, and other business data, one of the most promising areas for the use of big data is the use of existing Statistical and Administrative data. Data matching, at the micro or even sub-aggregate level, across statistical agencies and with non-statistical agencies can produce large gains at a relatively low price. Example include the use of medical and health insurance records for the construction of medical care price indexes, the matching of business registers, establishment, and enterprise data to produce for more detailed data on the characteristics of firms engaging in trade and foreign direct investment, and the use of motor vehicle registrations data to estimate

depreciation schedules. In general, expanded access to tax and administrative data for statistical purposes has the potential for quickly producing large benefits in the accuracy, level of detail, and efficiency of official statistics.

L. Conclusion

58. This paper has provided an overview of big data and their use in producing official statistics. Although advances in information technology, data sources, and methods have driven interest in the use of business and administrative government data collected and used for non-statistical purposes, use of such data is not new. Nor is it likely to be a panacea for statistical agencies confronting demands for more, better, and faster data with fewer resources. However, with careful attention to incentives; protection of privacy through data protocols and collaborative agreements; and integration of these non-statistical with existing statistical data, big data can play an important role in improving the accuracy, timeliness, and relevance of economic statistics at a lower cost than expanding existing data collections.

Appendix B: Standard Elements of A Model Agreement for the Provision of Administrative Records for Statistical Purposes (US OMB M-14-06, February 14, 2014)

1. Parties to the Agreement
2. Legal and Programmatic Authority
3. Duration or Period of Agreement
4. Purpose
5. Use of Data
6. Data Quality
7. Roles and Responsibilities for Data Protection
    a. Confidentiality and Privacy
    b. Data Security
    c. Data Transfer, Media and Methods for Transmission of Data
    d. Record Keeping, Retention, and Disposition of Records
8. Specific Penalties for Unauthorized Disclosure of Information
9. Potential Work Constraints
10. Breach
11. Disclaimers
12. Reporting
13. Administrative Points of Contact
14. Funding Information
15. Estimated Costs and Payment
16. Resolution of Conflicts
17. Modification/Amendment of Agreement
18. Cancellation of Agreement
19. Periodic Review of Agreement
20. Concurrence and Agency Signatory