United States Department of Agriculture

National Agricultural Statistics Service

Research and Development Division
Washington DC 20250

April 2019

On Producing Estimates for NASS's Quarterly Hogs and Pigs Report

Gavin Corral, Seth Riggins, Emilola Abayomi, Luca Sartore, Yijun Wei, Nell Sedransk, Cliff Spiegelman, Linda J. Young

Table of Contents

Chapter 1: Overview and Introduction to Hog Inventory Models

Nell Sedransk

1 Background and Motivation

There are six principal economic indicators for the status of the US agriculture economy; the national inventory of hogs is one of these. Hog inventory estimates are produced quarterly by the USDA National Agricultural Statistics Service (NASS) through a process that begins with a survey of hog production operations and concludes with a report of nine official statistics (for seven original variables and two computed quantities) for the nation and for the states.

1.1 Motivation for Modeling Hog Inventory

Statistical modeling of the US inventory of hogs was not part of the original NASS response to the requirement to produce quarterly estimates for national and state totals. Originally, the production of hog inventory estimates started with a complex survey, which can be briefly summarized as a stratified design within each state, based on operation size, i.e., total hog inventory as reported in another survey. Standard sampling estimators were used to calculate weighted state-level estimates ("preliminary state estimates") that were then compiled into preliminary national estimates. All these preliminary estimates were provided to a group of USDA livestock statisticians (Agriculture Statistics Board or ASB) who met to set the *official estimates* (national and state) after taking into account the preliminary estimates and other information available to them. Thus there was no basis for estimating standard errors or other measures of uncertainty associated with the Board's official numbers.

Modeling is the natural avenue for providing statistically sound and efficient estimates with statements of uncertainty. A comprehensive picture is essential to model formulation. In this case the goal is to meet time and accuracy requirements with acceptable uncertainties for hog inventory estimates (seven variables plus specified sums of some subsets) given the available data resources, both historical and current.

This larger picture for modeling hog inventories also has a time dimension with relationships across variables and reporting periods; hence estimates must be consistent across time and must reflect hog biology. In addition, a model must be (promptly) sensitive to shifts, trends and other shocks. Ultimately the model's accuracy, precision and timeliness depend directly on the available data and on computational feasibility.

This chapter outlines the properties desired in a hog inventory model, identifies the available data (both current and historical), considers the necessary elements in the model and the constraints and then outlines the challenges that make this modeling task unique.

1.2 Salient Facts about Hog Production

Hogs are raised for market in all 50 states; but 16 major hog-producing states account for a great majority of hogs produced (>95%); and another 14 states produce most of the rest. Thus national figures are driven by the 16 major hog-producing states although in other states as well, hogs contribute to the state's agricultural economy. Hog production operations can be as small as a few hogs, but there are operations of all sizes up to million-hogs-per-year operations. And even these largest operations ("extreme operators") may be only individual parts of a mega-corporate entity. The largest number of operations are small in size; nonetheless the national inventory numbers are driven by a relatively small number of very large operations (for the hog survey, these that are sampled with probability one).

These very large production operations, like production operations in other arenas, function in a highly regular manner that is assisted by breeding highly uniform hogs with nearly constant litter rates and a virtually known rate of growth. Thus, absent an unanticipated disruption, these operations tend to stabilize the national production level while following a predictable slow increasing trend over years.

There are differences among these "extreme operations" as some are vertically integrated from breeding sows to raising and marketing hogs; other operations are partitioned into breeding operations and feeding/finishing operations. From a national perspective, the mix of operation types is not important; from the state level it may become an added issue for some states.

Small operations function differently; many adopt a traditional model (breeding through marketing) with slightly lower litter rates than the extreme operations. But others are feeding/finishing operations.

2 Model Properties, Constraints and Challenges

To be valuable in the process of generating quarterly official estimates, a hog inventory model must not only reflect the information in available current data, but also reflect the inherent biological constraints. Estimates must be accurate with acceptable levels of uncertainty and, at the same time, meet logical conditions over time – it is not possible to have more large hogs in a quarter than there were small hogs previously just as it is not possible to slaughter more hogs than have been raised.

The first focus for modeling the hog inventory has been on the national inventory. However, the intention has always been to construct a model that could be expanded to provide direct estimates at the state level.

2.1 Under Equilibrium Conditions

Under equilibrium conditions, the national inventory changes predictably from quarter to quarter following an annual cycle that reflects seasonal change (weather, cost of feed, market forces, etc.) and overall exhibits a slow trend of expansion over time. Regionally the seasonal change might vary, but the stability of extreme operations controls the magnitude of change. Consequently for a single quarter, a model that incorporates both cyclic changes and longer-term trend can produce reasonable national estimates.

The logical constraints (under equilibrium) are effectively first, the interdependencies of the data within and across quarters, and second the relationship to the "gold standard" of (national) slaughter numbers.

Data interdependencies arise from the hog biology and growth pattern. Hog growth can be modeled as a function of time with little variation from birth to weaned to growth to a narrow range of ideal market weight. Survival post weaning can be modeled as well, noting that death loss is rare once a hog reaches a certain weight. Consequently the progression of each cohort of hogs through the four weight groups as a function of time allows prediction of the inventories for the weight groups at any point in time, provided the date of farrowing and the size of each cohort are known.

2.2 Under Dynamic Conditions Due to Disruption

Disruptions that create dynamic conditions include disease epidemics, natural disasters, expansion or contraction of slaughter house capacity, and, at least potentially, economic factors. The conditions created by each of these disruptions have different features. Disease epidemics, depending on the disease and on the operation's response to an outbreak, may result in a reduced litter rate (pre-weaned loss of piglets) or may predominantly affect the pig crop and/or the smallest hogs. The spread of an epidemic almost always has a spatial component. A natural disaster may be localized but affect all weight groups. Change in slaughterhouse capacity may prompt changes in numbers of sows bred and farrowed, hence affecting the size of the pig crop but not the litter rate.

The challenge is two-fold: detection and model adaptation. Detection may come from divergence of data from prediction or failure of data to conform to (logical) constraints. A

model may accommodate disruption in any of several ways. For example, the model may directly incorporate terms or components to accommodate disruptions. Alternatively, diagnostics might signal switching among multiple model versions or using these to create a mixture.

2.3 Differences for State-level Estimates

Patterns seen with national inventory numbers do not in general hold for state inventory numbers, which are more volatile. Since shocks tend to be local at any point in time (e.g., outbreak of disease or natural disaster) or possibly regional (expansion of slaughterhouse capacity in a saturated region), the proportion of affected operations is greater for a state than for the nation. The shock may spread (epidemics); also operators' responses may differ based on operation size. Thus at the state level, for a mix of sizes of operations, shocks escalate the modeling challenge to integrate the historical pattern (time series with constraints) with a more sensitive biologically-based prediction model. Operation information (monthly) on sows farrowed and pig crop (post-weaning) provides a different basis for prediction of inventory at any later point in time by projecting the survival and growth rates for monthly hog cohorts. Deviation from such predicted weight class inventories present a second source of diagnostics for detection of shocks and consequent predictions.

3 Information Needed – Information Available

3.1 Current Survey Data

Current quarterly survey data (either operation level or aggregated with adjustments for non-response) is needed for all (seven) variables and (two) calculated quantities together with indicators for imputation, operation size, and operation type. More frequent information on breeding, farrowing, pig crop (litter rate) is needed to define cohorts. To incorporate a spatial component, data must be localized for both the sampled and the non-sampled farms.

Available: Quarterly survey data aggregated (sampling estimators) to give state totals

Quarterly survey data broken down by month of quarter for sows farrowed, pig crop

Operation level data is made available <u>subsequently</u> but not in time for use in the current model (and has been used extensively in model testing and evaluation)

3.2 Historical Information

Past data and past official estimates are required for use in modeling historical patterns and trends. *Initial official estimates* are made (as described below) by the ASB in the quarter when the data are collected. These estimates are revised quarterly until the fourth revision (one-year post data-collection) which is the official *final estimate*. As is true for other federal data sources, the revisions allow information that is subsequently available to be taken into account;

in the case of the hog inventory, the additional information includes final slaughter numbers, and inventory estimates from subsequent quarters as well as external information such as the extent of a disease outbreak. Thus the best measure of model accuracy will be consistency with the official *final estimates*.

Available: Complete data base for 2008-2017 with no changes of definitions for:
initial through *final estimates* (seven variables, two calculated quantities)
Operation level reports for 2008-2017

3.3 Biologic Understanding

Growth and survival functions under normal and the various disturbance (disease) conditions

Available: Reference materials (see later sections)

3.4 Disturbances

Occurrences by location and date and duration (if appropriate) are needed together with the required responses. For natural disasters, maps that can identify both sampled and non-sample operations that are affected.

Available: Operation-level reports for 2008-2017
Repositories for disease reporting

4 Role of Model in Current Process

The current process calls for providing the ASB with model-based estimates of the national inventory (for nine quantities including seven variables and two quantities calculated from them) together with the standard errors for those estimates.

At the state level, data are collected, cleaned and aggregated into state totals. This part of the process includes imputation and adjustment for non-response. A separate project at NASS is considering these issues, so they are not part of the hog inventory modeling project.

A much-simplified depiction of the current process in Figure 1 shows how model-based estimation now fits into the process for setting the official estimates.

State recommendations are provided separately for a single sum (pig crop plus all four weight groups). The first task in preparing the data for incorporation into the model is to adjust the

sampling estimated totals to conform to the state recommendations (without violating the logical constraints imposed by past quarters' data).

The model then provides preliminary estimates for all seven variables and two calculated quantities. These are delivered to a "Pre-Board Panel" consisting of livestock statisticians at NASS Headquarters. This Pre-Board Panel takes into account data from other sources and other compilations from the survey and returns adjusted numbers to be incorporated into the model for a second run. The model estimates from the second run are delivered to the full ASB for their deliberations after which the ASB sets the *final estimates* for publication.

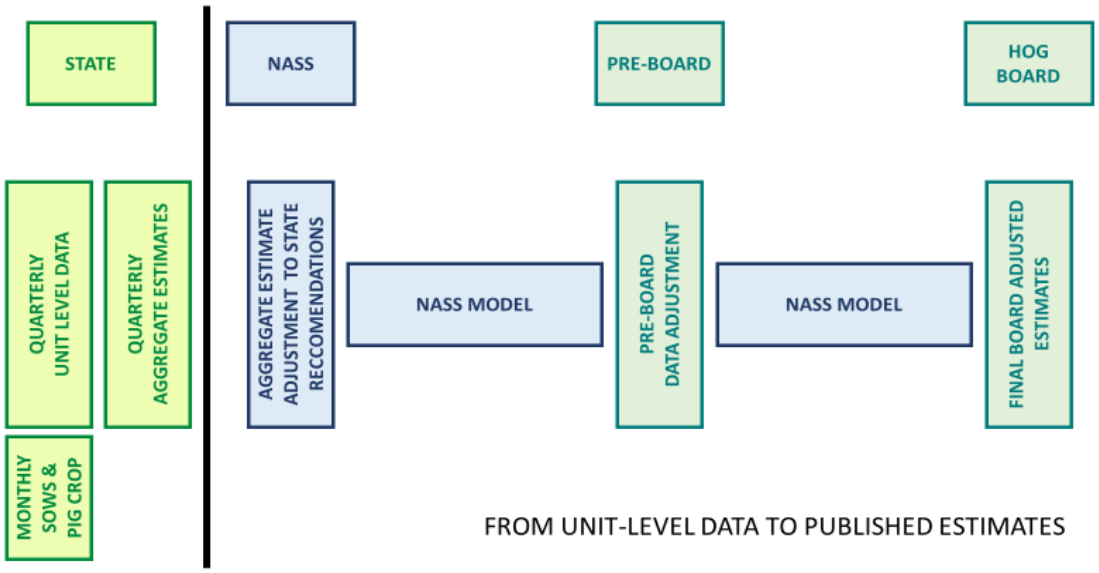

Figure 1. Estimating Hog Inventory: From Survey to Official Estimate

5 Possible Modeling Structures

5.1 Challenges

The critical problems in formulating an equilibrium model are to combine historically based patterns (annual cycle and slow trend) with a biologic model (or constraints) of hog growth and survival, consistency from one quarter to the next, and conformance to an external "gold standard."

The additional problems in extending from an equilibrium to a dynamic model are to introduce a spatial component and to project the dynamics of the time course of the disturbance. Technical issues also are challenging in determining how to deal with non-sample operations at risk for disease, for example, whether or not some of the sampled units are affected.

Thus far, work has focused on estimation of the national hog inventory. Once this has been done satisfactorily, attention will be turned to state-level estimates. Several different approaches have been proposed previously; two have been implemented without complete success.

5.2 Time Series Models

5.2.1 Busselberg embedded strict constraints in a frequentist state-space model (see Chapter 4).

5.2.2 A Bayesian time series approach was formulated to relieve the rigidity of the Busselberg model, taking advantage of the prior distribution to mix (flexibly) an equilibrium model with a shifted model for disturbance

5.2.3 A superposed time series model would allow multiple secondary functions for potential alternative effects of disturbances

5.3.4 A variant of 5.2.3 would introduce a "switch function" to select the specific time series model when each of the possible models was developed from the equilibrium model.

5.3 Multivariate Models

5.3.1 An unconstrained multivariate model with optimal length time window for inclusion of past data (see Chapter 4)

5.3.1 An independent spatial model estimating extent and strength of disturbance effect, used to adjust equilibrium model.

6 Conclusion

Formulating a hog model that will handle all the complexities is difficult. It is not clear that all possibilities should be built into a single model. Regardless, along the way it may be equally important to establish diagnostics that can point to the dynamics (or not) that are evident in the data and hopefully are confirmable from external information as well as expert opinion.

Chapter 2: Quarterly Hogs and Pigs Reports

Seth Riggins

The USDA National Agricultural Statistics Service (NASS) produces six of the nation's principal economic indicators, major statistical series that describe the current condition of the nation's economy. One of these is the quarterly Hogs and Pigs Report (see the appendix for the March 2019 and December 2018 reports). The report is produced quarterly in December, March, June, and September, and includes the following official statistics:

1. **Breeding Herd** - Includes boars; sows; gilts and young males kept for breeding.
2. **Market Hog Inventory** - All hogs and pigs intended for market.
3. **Weight Groups** - There are four weight groups (less than 50 lbs., 50-119 lbs., 120-179 lbs. and 180+ lbs.) for the Market Hog Inventory.
4. **Farrowings** - The number of sows that farrowed (gave birth) during a given time period.
5. **Pig Crop** - The number of live birth pigs that were alive and still owned by the operation, sold, or slaughtered by the reference date on the questionnaire (pigs that died before the reference date are not counted for the purpose of the pig crop, though if they were weaned before death they are included in the death loss estimate).
6. **Litter Rate (or Pigs Per Litter)** - Equal to Pig Crop divided by Farrowings for a given time period.

Sixteen states account for more than 95% of the hog and pig production: CO, IL, IN, IA, KS, MI, MN, MO, NE, NC, OH, OK, PA, SD, TX, and UT. In addition to the national estimates, estimates are produced for each of these states in March, June, and September. In December, estimates are produced for all states and for the nation.

NASS has two sources of data for producing the report: the NASS Hog Inventory Survey and administrative slaughter data. The Hog Inventory Survey is conducted quarterly in December, March, June, and September. The survey's target population is all hog and pig producers in the United States. The reference date is the first day of the report month, that is, December 1, March 1, June 1, and September 1. The data collection period is 20 days. The sample size for December 2018 was 8500, of these operations 6100 or 71% responded. In March 2019, the participation of fewer states led to a reduced sample size of 5100 with 3500 (69.2%) respondents. (A more complete description of the survey process is in Chapter 3.)

After the data collection period ends, NASS Regional Field Offices have about four or five business days to complete editing and analysis, execute the summary, and interpret the survey results. Regional Field Offices are responsible for performing a detailed review of their survey

results. Any irregularities revealed by the summary must be investigated and, if necessary, resolved. Using the historical relationship of the survey estimates to the official estimate, Regional Field Offices interpret the survey results and submit a recommended estimate to

Headquarters (HQ) for all data series for which their region is in the NASS program. That is, both the survey estimate and the state recommended estimates based on the Regional Field Office review are submitted to HQ in Washington DC. The historical relationship of the survey estimates to the official estimate over time is evaluated using tables and graphs to determine accuracy and bias.

Slaughter data are the only administrative source of information for hogs and pigs. These data are provided to NASS by the inspectors of USDA's Food Safety and Inspection Service (FSIS), who collect data and demographic information on the regulated slaughter facilities. The number of pork carcasses is enumerated and can be combined with other datasets to enhance the analysis. These data are available on a weekly basis and consist of:

- several variables describing the establishments that process meat, poultry and eggs,
- the inspection activities,
- the slaughter variables and other information about the products and their safety for human consumption.

Once state-level estimates are submitted to HQ, the HQ Hog Statistician reviews the state-level survey estimates, the state recommendations, and regional changes. The state-level comments are reviewed to determine whether any large outbreaks of disease were reported and whether any unusual weather impacted the hog industry or survey process in each state. The state-level estimates are then compiled by the HQ Hog Statistician to give national estimates. Within eight working hours of receiving the estimates, the HQ Hog Statistician meets with the Livestock Section Head, Livestock Branch Chief and the Methods branch person who ran the edit and summarization procedures. As a group they review the estimates and administrative data and compose two or more scenarios for the Agricultural Statistics Board to review the following day. Each scenario integrates the state-level interpretations of the data to provide an interpretation of the national estimates. As an example, an unusual decrease in the estimated piglet crop may be attributed to a disease outbreak in some states. An increase in the number of sows farrowed may be explained as a response of producers to a disease outbreak or a change in market conditions.

NASS employs a balance sheet approach as part of the evaluation of estimates. The balance sheet reflects the biological constraints of production. As an example, since the time required

from birth of a piglet to slaughter is about six months, it is not possible to slaughter more hogs than there were piglets six months earlier. The supply components of the United States balance sheet are the beginning inventory, births, and imports (in-shipments for State balance sheets).

From this supply, the disposition components – commercial slaughter (marketings at State level), farm slaughter, deaths, and exports – are subtracted. The result is the estimated number on hand at the end of the period or year. Commercial slaughter is an important element of the balance sheet at the national level since its high degree of reliability is based on a near-actual count of animals slaughtered. With respect to modeling, balance sheets will be more fully discussed in Chapter 4. An expanded write up of Balance Sheets as they pertain to the estimation process is included in the Appendix.

Methods Division staff provide ratio reports for the top 100 largest producers, combining the total operations for each producer. The ratio reports are for current quarter to previous quarter and current year to previous year. These provide a snapshot of what the largest group of hog producers have in inventory, farrowings and pig crop. This is supplemental information derived from the survey respondents and is not adjusted for non-response.

A constrained state-space model that incorporates the NASS survey estimates, the slaughter data, and accounting constraints arising from biological considerations produces estimates of the quarterly total hog inventory, pig crop, farrowings, and litter rate (Busselberg, 2013). Since an extended Kalman Filter is used to integrate the disparate data, this model is referred to as the KFM. This model is described in more detail in Chapter 4.

When the industry is not experiencing a shock, such as a major disease outbreak or natural disaster, the above information is generally sufficient to produce accurate official statistics. However, especially in the early stages of a shock, it is challenging to produce accurate estimates. To aid in identifying that the industry is in the early stages of a shock, a series of diagnostic plots from Bayesian Hidden Markov Models are developed (Wang et al. 2016). These plots are described in more detail in Chapter 4.

To give an idea of the time available for modeling, the HQ Hog Statistician completes his work and enters data into the system about 1:30 pm on a Wednesday, for example. The KFM model results and diagnostic plots are due by 8:30 am the next morning, and the Hog Statistician loads them into the NASS system.

Before the Agriculture Statistics Board (ASB) meets to discuss the national estimates on Friday morning at 8:00 a.m., a pre-board meeting is held at 9:30am on Thursday. The Hog Statistician,

the Livestock Branch Chief, the Livestock Section Head and a representative from Methods Branch are always on the pre-board panel. In the pre-board meeting, the pre-board panel members review all available information (survey estimates, slaughter data, balance sheet numbers, state recommendations, ratios of current year and quarter to previous year and quarter for top producers, the KFM estimates, and diagnostic plots of the presence of a shock) and establishes preliminary national estimates for the current quarter.

When the panel adjourns from the pre-board, the preliminary estimates and revisions are loaded and made available for modeling. This process is usually completed between 1 and 2 pm that same afternoon (Thursday). The KFM is rerun using the same information as for the pre-board AND the preliminary estimates and revisions from the pre-board. The results from the final KFM run is due by 4 pm that same afternoon.

The Head of the ASB, the Statistics Division Director, the HQ Hog Statistician, representatives from Methods Branch and Survey Administration Branch, and two or three Regional Field Office personnel from major hog states (depending on the quarter) are on the ASB for hogs and pigs.

The ASB convenes the following morning (Friday) at 8:00 am to review all information and estimates. NASS employs the "top-down" approach by determining the national estimates first and then reconciling the state estimates to the national number for each published estimate of the hog inventory, pig crop, and farrowings. In addition, the official estimates from the previous three quarters are reviewed in March, June and September quarters. The previous seven quarters are reviewed in December quarters. These may be revised based on the slaughter data that have been collected during the quarter since the last report or updated information from hog operators. The largest changes in the official statistics usually occur during the process of the first revision. The change may, for example, reflect the impact of a weather event that had either a larger or a smaller effect than thought at the time of the original board. Minimal revisions are usually made to the official statistics being reviewed for the second or third time.

After three revisions, the official statistics become "final". Every five years after a Census of Agriculture is conducted, the previous twenty quarters are open for review and revision. This is the final time in which revisions may be made.

Over the next week, state level estimates are reviewed and revised in order to meet national-level targets. Approximately seven days after the ASB meets, an executive briefing is given while "locked-up" to the Secretary of Agriculture (or his designee) around 15 minutes before the official release time. The lock-up procedure ensures that no communication with the outside world is permitted while the report is presented (see Allen (2007) in the appendix for a

more complete description of lock-up). Part of the executive briefing is a discussion of pre-report trade expectations from non-government sources. These pre-report expectations generally focus on percentage changes from previous year and provide a snapshot of what the industry at large expects from the report. The pre-report trade expectations are only included in the briefing so the Secretary of Agriculture is informed with what the markets are forecasting. The ASB does not see the trade expectations until the report is finalized and the trade forecast has no bearing on any deliberations. The official statistics are then released in the form of the Hogs and Pigs Report, usually at 3:00 p.m. The estimates are generally released to the public by the last week of the month. The publication date and time may change due to the timing of federal holidays.

Every five years NASS conducts the Census of Agriculture, which enumerates all known farms and ranches across the United States. The information gathered from the Census of Agriculture is used to establish "bench mark" levels by which the survey estimates can be compared and bias determined. Survey-based estimates can also be impacted by outliers – individual reports that have excessive influence on the results due to either improper classification or extremely unusual data for a given operation (i.e. the operation is not representative of other operations).

NASS thoroughly reviews the survey data to identify these situations and considers their impact on the survey results when establishing the official estimates.

Work Cited

Allen R (2007). Chapter 5: Current Practices and Procedures. *Safeguarding America's Agricultural Statistics: A Century of Successful and Secure Procedures*. Washington DC: USDA NASS. Pp 22-26. Available at https://www.nass.usda.gov/About_NASS/pdf/asb_historical.pdf.

USDA National Agriculture Statistics Service (2018). *Quarterly Hogs and Pigs.* Available at https://downloads.usda.library.cornell.edu/usda-esmis/files/rj430453j/bc386p647/rf55zc904/hgpg1218.pdf.

USDA National Agriculture Statistics Service (2019). *Quarterly Hogs and Pigs.* Available at https://downloads.usda.library.cornell.edu/usda-esmis/files/rj430453j/k930c4962/ft848z158/hgpg0319.pdf.

Chapter 3: The Hog Inventory Survey

Emilola J. Abayomi

1. Introduction

The quarterly Hog Inventory Survey (often called the hog survey) provides the primary data and subsequent estimates used by the Agricultural Statistics Board to develop the Hogs and Pigs Report, one of the six principal economic indicators produced by NASS. The target population for this survey is all US farm producers who own at least one hog or pig on the survey's reference date. At the end of 2017, the estimated number of farms producing at least one hog or pig was 66,439, and nationally the number of hogs and pigs was more than 72 million (2017 Census of Agriculture). As is the case more generally in agriculture, the number of producers with medium-sized operations decreased from 2012 to 2017, while the numbers of producers with small and large operations either increased or held relative steady as shown in Table 1. As a consequence, in 2017, farm operations with at least a thousand hogs and pigs accounted for 97% of the US hog and pig population, compared to just under 96% in 2012.

Table 1. Number of farms with specified range of hogs and pigs as of December 31, 2017

Farms with	2017	2012	Farms with	2017	2012
1 to 24	46,475	41,688	500 to 999	1,305	1,977
25 to 49	3,759	3,435	1,000 to 1,999	2,016	2,677
50 to 99	1,889	2,161	2,000 to 4,999	4,724	4,718
100 to 199	1,220	1,469	5,000 or more	3,600	3,006
200 to 499	1,451	2,115			

In this chapter, the Hog Inventory Survey process is fully described. The characteristics of the data and the factors affecting data quality are highlighted.

2. Sample Design

The hog survey is conducted quarterly, in December, March, June, and September. The reference date for the survey is the first day of the survey month. The survey results are combined with other information available to the ASB (see Chapter 2) to produce the quarterly Hogs and Pigs Report. This report is released by the end of the survey month with only rare exceptions, which are usually due to federal holidays.

The primary variables to be estimated by the hog survey are total inventory, breeding herd (boars, sows, gilts and young males kept for breeding), market hog inventory, four weight groups (less than 50 lbs., 50 to 119 lbs., 120 to 179 lbs., and at least 180 lbs.), farrowings, pig crop, and litter rate). The number of sows expected to farrow in 4 to 6 months and the number expected in 1 to 3 months (the intentions) are also estimated. The number of sows expected to farrow in 1 to 3 months is a revision to the intentions reported for 4 to 6 months the previous quarter. In addition to the quarterly totals for these 9 variables, operations supply that quarter's monthly breakdown for sow farrowings and pig crop. Since the focus is on hog and pig production, the importance of obtaining information from the large producers is apparent. In December, official statistics are provided for all states. In March, June and September, state estimates are provided for only the 16 major hog and pig producing states, i.e., those states with the largest hog and pig production (CO, IL, IN, IA, KS, MI, MN, MO, NE, NC, OH, OK, PA, SD, TX, and UT).

NASS uses a dual frame approach, consisting of the Hog Survey list frame (a list of all known US agricultural operations with at least one hog or pig owned) and the NASS area frame. The Hog Survey list frame is created from the NASS list frame, which includes all known farms in the US. It includes all operations with hogs and pigs except those for which the operation has less than 500 hogs and the control data precede 2007. The frame accounts for about 97% of all hog and pig production. The June Area Survey, which is drawn from the area frame, is used to adjust for the 3% undercoverage of the list frame. The sample size drawn from the Hog Survey list frame in December 2018 was 7,589; it was 4,899 in March 2019.

The sample drawn from the list frame has a hierarchical stratified random design. The objective is to achieve a 1% CV for Total Hogs and Pigs Inventory at the national level, a 3% CV for the 7 critical major hog and pig producing states (IL, IN, IA, MN, MO, NE, NC), 6% CV for the remaining hog and pig producing states for which official statistics are reported, and 6% for the estimate for the annual states combined. The sample size required to achieve these targets is adjusted upwards to account for anticipated nonresponse when setting the survey's sample size each quarter.

In December, all states are sampled. In March, June, and September, samples drawn from the 16 largest hog and pig producing states are large enough to meet the CV targets for those states. Smaller sized samples are drawn from 14 additional states that have substantial hog and pig production (AL, AR, AZ, CA, GA, KY, MS, MT, ND, SC, TN, VA, WI, and WY). Official state estimates are not published for these 14 states in March, June, or September; however, an aggregate estimate is published as Other States. The remaining states (AK, CT, DE, FL, HI, ID, LA,

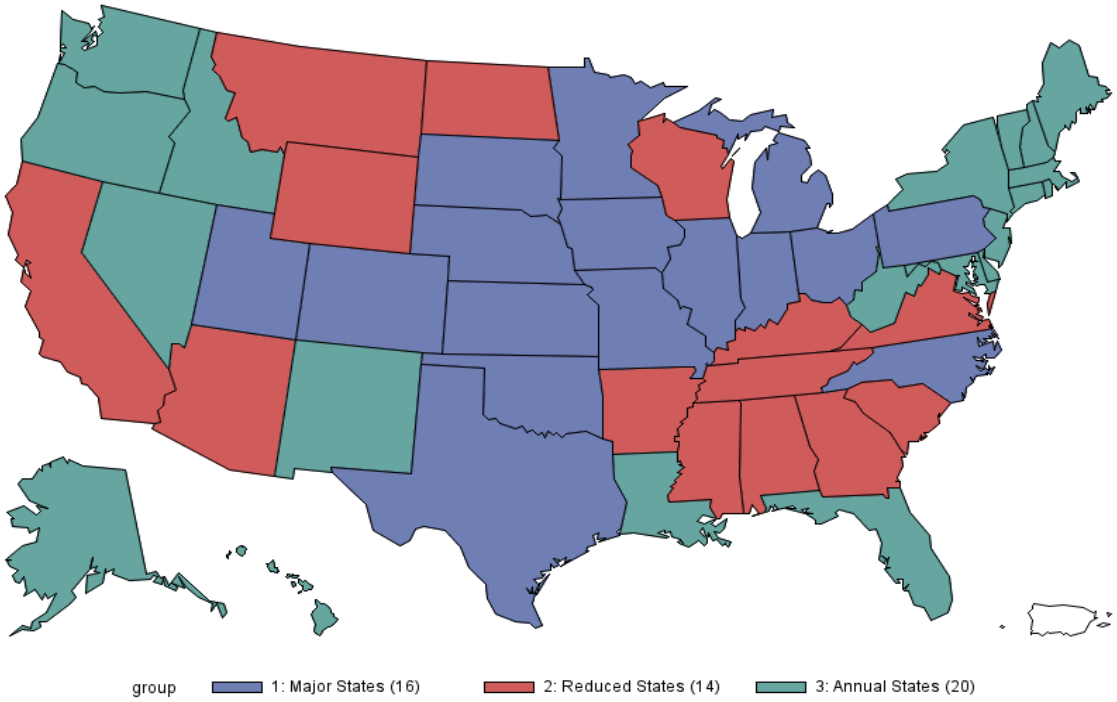

Figure 1. States for which state estimates are published in all quarters (red), states that are sampled in all quarters but for which state estimates are published only in December (blue), and states that are only sampled and published in December.

ME, MD, MA, NV, NH, NJ, NM, NY, OR, RI, VT, WA, and WV) are only sampled in December. See Figure 1.

A stratified random sample is drawn from each sampled state. The strata are defined by the total number of hogs and pigs owned. The control data on the list frame, which consist of information about each operation obtained through earlier censuses or surveys, are used to determine stratum boundaries. Because the distribution of operation sizes differs from state to state, stratum specifications vary with state. Operations in one or more of the state's top producing strata, are sampled with probability one. The reciprocal of the probability of sample inclusion, which is constant for all records in the same stratum, is the sampling weight for each record drawn from the stratum. See Table 2 for the strata definitions and sampling weights for Iowa in 2018. The appendix has the same information for other states. Beginning in December 2019, an extreme operator sow stratum will be added for Iowa, Minnesota and Nebraska. Units in this stratum will be selected with probability one. The sample size in the stratum of the smallest producers (stratum 80) will be reduced substantially for all states.

To assess undercoverage for NASS surveys, the NASS list frame is matched to the JAS records using probabilistic record linkage. All JAS records that do not match are said to be Not-on-List (NOL) records. The subset of the NOL records that had at least one hog or pig owned, termed the NOL sample, is used for estimating undercoverage of the Hog Survey list frame. The NOL sample in December 2018 was 738.

Table 2: Iowa Strata for the Hog Inventory Survey		
Stratum	Number of hogs and pigs	Weights
80	1-99	24.00
82	100-999	2.19
86	1000-9999	1.53
88	10000-29999	1.00
90	30000-49999	1.00
92	50000-89999	1.00
98	90000+	1.00

A new sample is drawn for the December survey each year. As a panel survey, producers selected for the sample are asked to respond in all quarters for which their state is included in the sample (all states not in green in Figure 1). In December, data are collected for all states and the NOL records that had at least one hog or pig, had positive or unknown hog intentions, or had previously owned hogs. In March, June, and September, data are collected for the sample in the 16 largest producing states and in the 14 reduced sample states. Data for the NOL records and remaining states are modeled in March, June, and September.

All federal data collections require approval by the Office of Management and Budget (OMB). NASS must document the public need for the data, apply sound statistical practice, prove the data does not already exist elsewhere, and ensure the public is not excessively burdened. The Hog Survey questionnaire must display an active OMB number that gives NASS the authority to conduct the survey, a statement of the purpose of the survey and the use of the data being collected, a response burden statement that gives an estimate of the time required to complete the form, a confidentiality statement that the respondent's information will only be used for statistical purposes in combination with other producers, and a statement saying that response to the survey is voluntary and not required by law.

3. Data Collection

For consistency across modes, the paper version is considered the master questionnaire and the web and Computer Assisted Telephone Interview (CATI) instruments are built to model the paper instrument. Questionnaire content and format are evaluated annually through a specifications process in which requests for changes are evaluated and approved or disapproved. Input may vary from question wording or formatting to a program change involving the deletion or modification of current questions or addition of new ones. If there are substantive changes to either the content or format proposed, a NASS survey methodologist pre-tests the changes for usability. Prior to the start of data collection, all modes of instruments are reviewed and web and CATI instruments are thoroughly tested.

Sampled farms and ranches receive a pre-survey letter explaining the survey and informing them that they will be contacted for survey purposes only. The letter provides the questions to be asked to allow respondents to prepare in advance and also provides a pass code they can use to complete the survey on the internet. All modes of data collection (web, mail, telephone and in-person) are utilized for hog surveys. Regional Field Offices are given the option of conducting a mail out/mail back phase. Although mail is the least costly mode of collection, the short data collection period (20 days beginning on the quarter's reference date) and the uncertainty of postal delivery times limit its effectiveness. Most of the data are collected by CATI by the Regional Field Offices and Data Collection Centers. Limited personal interviewing is done, generally for large operations or those with special handling arrangements. A program is run to determine whether any sampled farms are in multiple on-going surveys, so data collection can be coordinated.

4. Editing and Estimation

Responses are required for the extreme operators, those sampled with probability one in the unbounded stratum (stratum 98 in Table 2). If an extreme operator refuses to respond or is inaccessible during the data collection period, the Regional Field Office staff manually imputes for the record. This manual imputation is generally based on previously reported data for that operation, perhaps adjusted for fluctuations from quarter to quarter. An interactive data analysis tool available to staff provides matched records ratios (measures of change) that are considered when manually imputing for an extreme operator.

The responses, including the data for the manually imputed extreme operators, are edited for consistency and reasonableness using automated systems. The edit logic ensures the coding of

NASS administrative data, such as response codes, reporting codes, and section completion codes, follows the methodological rules associated with the survey design. For example, if it is determined that an operation still has hogs and/or pigs, the section completion code is a 1; otherwise, it is a 2. Relationships between items on the current survey are verified and in certain situations item-level data in a current survey may be compared to item-level data from earlier surveys to make sure certain relationships are logical. (An item-level datum is the response to a question on the survey.) The edit determines the status of each record to be either "dirty" or "clean". Dirty records must be updated and reedited or certified by an analyst to be clean. If updates are needed, they are reedited interactively. Only clean records are eligible for analysis and summary.

In the analysis, the usable reports (those for which the response was complete, manually imputed or machine edited) are treated the same. For the extreme operations in the unbounded stratum, the survey weight is the sampling weight. For other records, the survey weight is the sampling weight adjusted for the nonresponse in the corresponding stratum. Two approaches are used to adjust the sampling weight for nonresponse. The reweighted adjustment is the reciprocal of the proportion of all usable reports within the stratum. The adjusted nonresponse weight adjustment uses an additional piece of information. When a sampled farm refuses to cooperate, interviewers will probe to determine the presence of hogs and/or pigs even though the number is not known. If it is found that hogs and pigs are present, the section completion code is 1. As the proportion of nonrespondents with a section completion code of 1 increases, the adjustment for nonresponse increases. The reweighted estimator is the design-based estimator that uses the reweighted adjustment for nonresponse. The adjusted estimator is the design-based estimator using the adjusted approach to adjusting for nonresponse. Each is used to obtain stratum and state estimates. Typically, the adjusted estimate for total inventory is 2 to 3% higher than the reweighted estimate. Both reweighted and adjusted estimates are provided to the ASB (see the Appendix for a fuller report of these estimators).

5. Data Considerations

A closer of look at the survey data provides insights into the relationship of the design-based survey estimates relative to the state-recommended estimates, the initial board estimates, and final board estimates. Given that the fully revised final estimates are the gold standard, the design-based survey estimates are biased downwards. The bias is real, i.e., the fully revised final official estimates are not biased upwards, because the final estimates are revised to be consistent with administrative slaughter data.

The survey data are also evaluated for early signs of the onset of a shock. The emergence of the porcine epidemic diarrhea virus (PEDv) in 2013 affected the hog population, making it challenging to accurately estimate total inventory. It is unclear whether the survey estimates detect the PEDv shock better or sooner than the state-recommended or initial board estimates. Towards the end of the epidemic in 2016, the survey estimates suggested an increase in inventory. The state-recommended and initial board estimates underestimated the increase but the revisions reflected in the final board estimate corrected for this initial under estimation.

Because original responses and imputed responses are treated equally in the analysis, it is natural to question whether imputed data impact the estimates. It should be noted that imputation occurs for all extreme operations with missing data since missing data are not allowed for these operations. For extreme operations in one state, the differences in hog totals between imputed data and reported data were compared from March 2010 to December 2017. A full report means the operation had a report for every quarter. Fifteen operations had reports for 31 quarters. On the plot, March 2010 is marked 1, June 2010 is marked 2, …, December 2017 is marked 31. The red symbol (A) denotes imputed data values and the blue symbol (B) denotes original data values. The mean imputed values were lower than the original values for most dates, but this is largely due to the fact that imputation occurred more frequently for the smaller of the extreme operations. The exceptions occur for dates 6, 24, and 29 (June 2011, March 2016, and June 2017, respectively), which are not epidemic years. There are no apparent indications that imputation impacts the estimates or dampens the effects of a shock during epidemic years. However, continued evaluation of imputation bias on estimates is needed.

Spread of an epidemic has a spatial component that could affect predictions, including "predicting the present" for non-sampled operations near affected operations. The rapid spread of porcine epidemic diarrhea virus (PEDv), a highly infectious disease, is a good example of transmission attributed to geographic proximity. PEDv causes outbreaks of acute diarrhea and vomiting in pigs and hogs. The disease is most severe in young pigs where mortality rates are high, resulting in a rapid decline in inventory. The virus was first documented in Ohio during May 2013. The US Department of Agriculture Animal Plant Health and Inspection Service (USDA APHIS 2014) issued a federal order in June 2014 requiring all hog operations, veterinarians, and laboratories to report any instances of the PEDv virus. Figures 3 and 4 are progression maps displaying the number of positive PEDv accessions by state in intervals prior to the federal order mandate (USDA APHIS, 2014). Because the identity of the hog operation from which the sample was taken was not provided to USDA with the laboratory testing results and because the likelihood of repeat testing on affected multiple herds, ***the number of infected herds within a state cannot be determined from these data***. However, they do provide insight into the spread

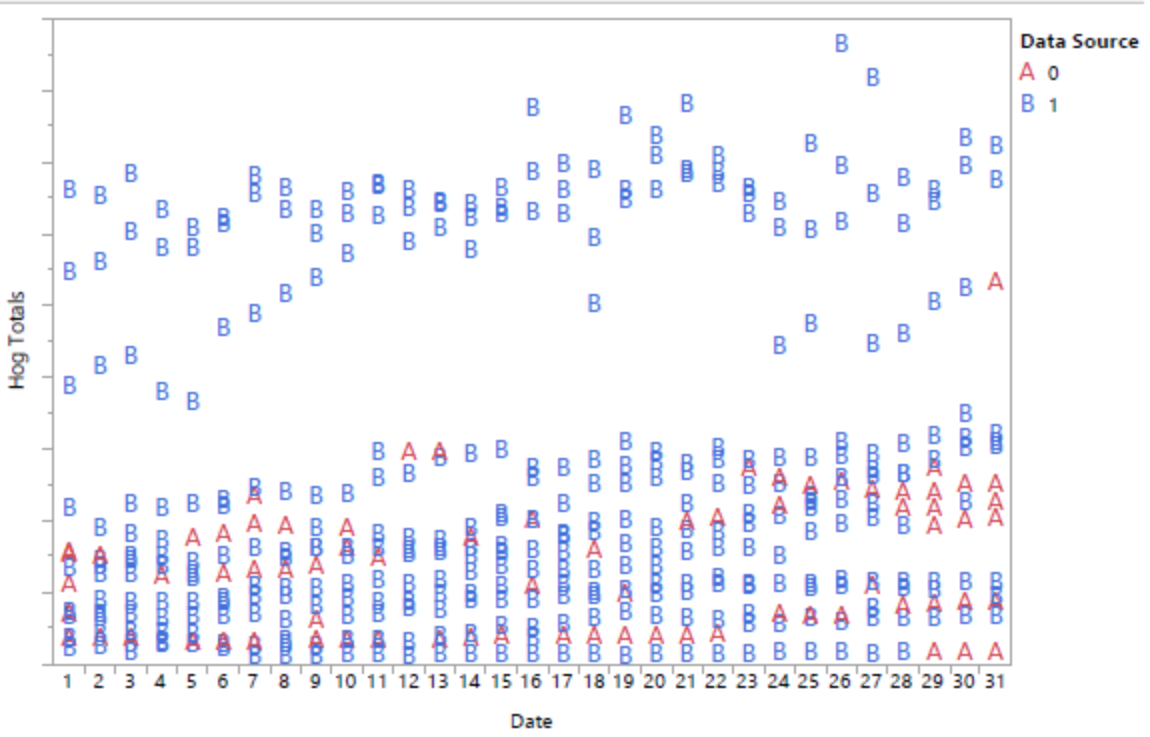

Figure 2: Plot of Imputed Data versus Original Data of Total Hogs Data Extreme Operations in Iowa from March 2010 to December 2017

of PEDv across states over time. The first map (Figure 3) dated July 1, 2013 shows nine states with positive cases of PEDv. The virus quickly spread to neighboring states. By March 1, 2014, over half of the states had positive cases of PEDv (Figure 4). More on PEDv testing can be found in the APHIS report in the appendix.

6 Work Cited

USDA-APHIS-VS Center for Epidemiology and Human Health (2014). Swine Enteric Coronavirus Disease Testing Summary Report. Available at https://www.aphis.usda.gov/animal_health/animal_dis_spec/swine/downloads/SECD_pre-fed_order_nahln_test_sum_rpt.pdf.

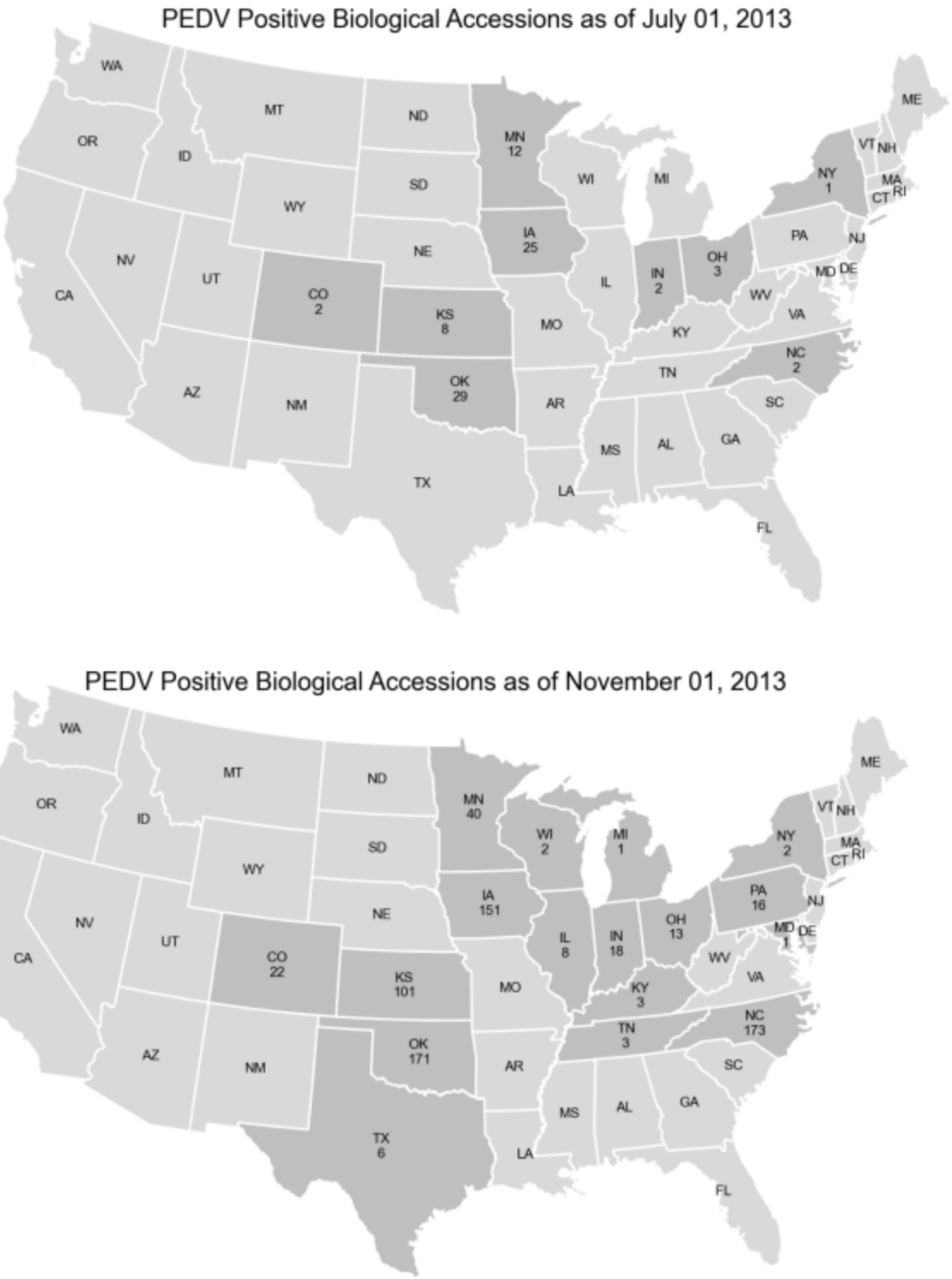

Figure 4: Maps of the number of positive PEDv accessions as of July 1, 2013 (above) and November 1, 2013 (below). From USDA-APHIS-VS Center for Epidemiology and Human Health, 2014.

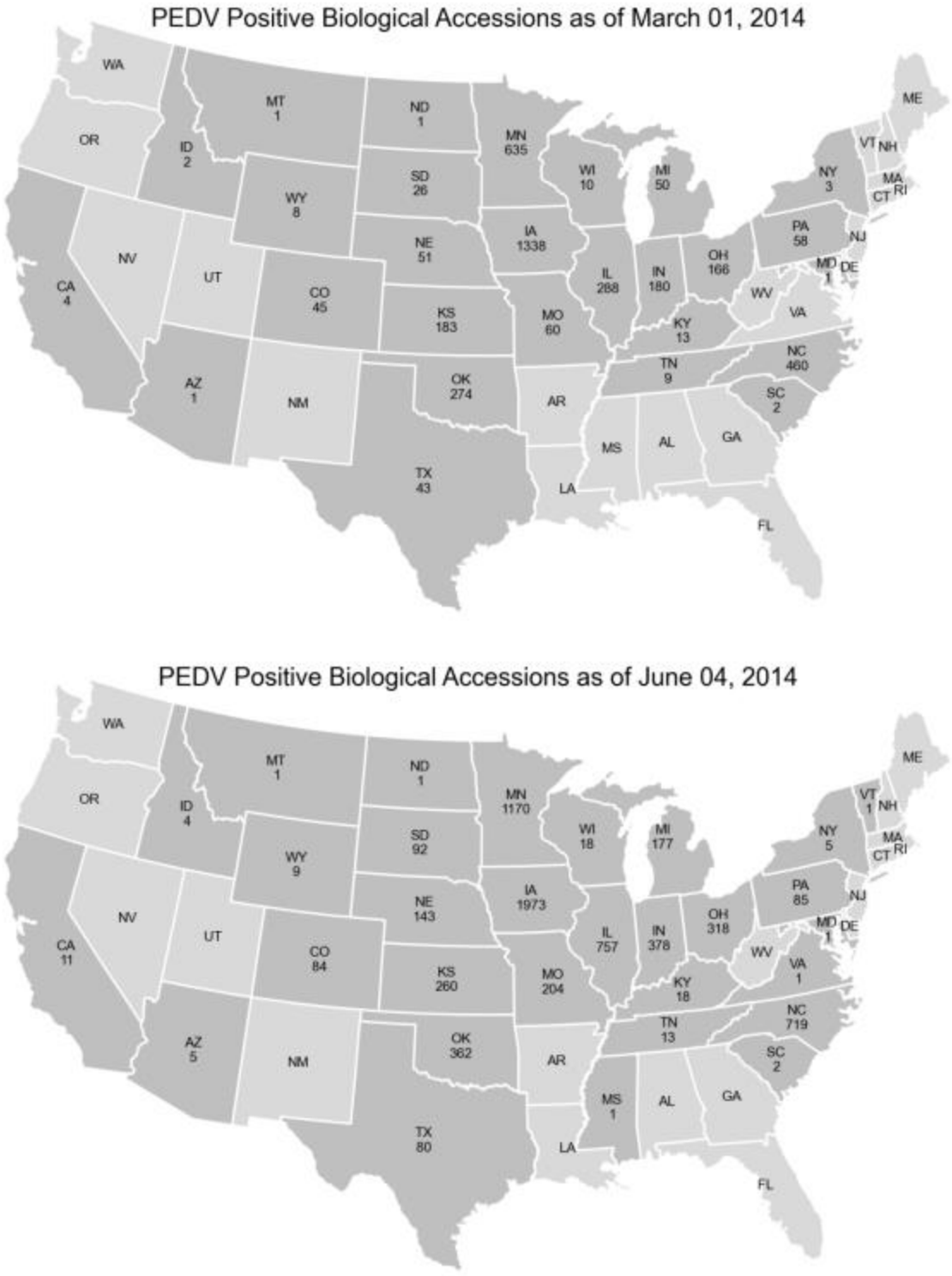

Figure 5: Maps of the number of positive PEDv accessions as of March 1, 2014 (above) and June 4, 2014 (below). From USDA-APHIS-VS Center for Epidemiology and Human Health, 2014.

Chapter 4: Modeling Efforts

Gavin Corral

1 Introduction

The purpose of this section is to identify the fundamental elements of a hog inventory model. Then two different modeling approaches developed at NASS are described and their performance evaluated. Analyzing the strengths and weaknesses of these two types of models lead to defining requirements for an improved model that can be employed at both national and state levels.

1.1 **Fundamentals for Modeling Hog Inventory**

The purpose of a hog inventory model is to provide estimates for the required total hog inventory and specified subpopulations that are coherent with respect interrelationships (constraints) and are efficient (as measured by CVs). In general, simply compiling the survey results fails in one way or another to satisfy the set of accounting relationships. These include, for example, relationships between current and past inventory, also the relation of external transaction data to current and past inventory, and while also reflecting accurately the hog growth cycle. Simply put, there cannot be more large hogs in a given quarter than there were small hogs in the previous quarter; also there cannot be more hogs slaughtered in a quarter than there were hogs large enough to be taken to market.

In addition to managing constraints, a model can also incorporate (annual) cyclical production patterns and overall industry trends, whether expansion or contraction. In periods of relative production/market equilibrium, these patterns and trends can be modeled well from historical data and updated with current survey data.

A particularly difficult challenge to estimating hog and pig inventory arises from the unpredictable occurrence of deviations from a pattern of relative equilibrium. Such shocks can greatly affect hog production, either locally or nationally depending on the nature of the shock. Shocks are defined here as events that cause hog and pig inventory to shift suddenly upwards or downwards. Examples of shocks include outbreaks of infectious diseases, natural or other disasters, sudden economic policies, or other disturbances that cause changes in hog inventory whether from the event itself or from the producer's response. Shocks, as in the case of epidemics, may have an immediate local effect but may then spread; or shocks may be universal in their impacts.

Therefore the modeling challenge is to develop a predictive model that: i) captures the equilibrium picture accurately, ii) detects and adjusts for shocks immediately when these appear, iii) accounts for the birth-to-market hog life cycle, allowing for disruption due to disease, disaster or other cause, and iv) satisfies external accounting relationships.

Two very different, complete working models have been implemented at NASS. Both of these models were developed to address the challenges of estimating the hog inventory, addressing the biological dynamics of the hog population and capturing the economic patterns including sudden departures from the equilibrium. Each of these models is successful in meeting some, but not all, of the four criteria above. Each is described below and its performance is evaluated.

The first model is a Kalman Filter model (KFM) for national inventory (only) originally developed by Busselberg (2013). This state space model is multi-dimensional with relationships among the estimated quantities embedded in the constraints imposed.

In a very different approach, a sequential general linear model (SGLM) was developed by Kedem and Pan (2016) to be sensitive to departures from the equilibrium pattern and to capture economic patterns affecting hog production.

1.2 Criteria for Model Evaluation

The stated tolerance for the official national estimate, for example for total market hogs, is plus/minus 500,000 hogs (one day's slaughter). In the quarter for which the estimate is published, the slaughter data are one benchmark.

To evaluate the model performance for all estimated quantities, model estimates can be compared with i) the official estimate issued by the Hog Board that same quarter, or ii) the final estimate issued by the Hog Board as revised four quarters later. In periods of relative equilibrium, there may be little difference between i) and ii). However with the occurrence of a shock, the severity of its impact may not be recognized immediately so that the initial official estimate will be revised at least once resulting in a sizeable difference between i) and ii). Comparing model estimates to i) measures contemporaneous agreement with the Hog Board. Comparing model estimates to ii) measures accuracy with respect to the best available information *a posteriori*.

Model estimates must also satisfy logical and accounting requirements. A biologic model for survival and growth of each (monthly) hog cohort, calculated for an equilibrium period, can provide expected hog inventories by weight class as a guide.

1.3 **Biological Considerations**

Disease outbreaks in the hog population are one of the primary causes of shocks. Between-farm direct or indirect contacts through transportation of animals or biological materials or cross-contamination through inputs, such as machinery or human workers, are among the most important factors to disease spread in food animals (Fèvre et al. 2006). Examples of diseases affecting the US hog industry that are spread due to farm-to-farm contact include porcine reproductive and respiratory syndrome (PRRS) and porcine epidemic diarrhea (PED). For both diseases, PRRS and PED animal movements (e.g. gilts, boars, weaned pigs, feeder pigs and cull animals) represent one of the most important disease transmission routes among farms (Valdes-Donoso et al. 2017). Furthermore, market responses to the anticipation or occurrence of disease outbreaks can cause upward shocks in inventory as producers try to stay ahead of disease outbreaks. Consequently, modeling efforts have tended to be overly constrained and unable to respond to the sudden changes of inventory or have not been able to stay true to the biological cycle of hogs.

Biological factors related to births, growth, health status, mortality risks, and disease exposure and spread affect the inventory. Explicit incorporation of these factors into a model should make the estimates and predictions of the targeted variables more reliable. That means that model-based estimates should be intrinsically connected to the flow of time and the weight gain. Hog production is a highly controlled process with hogs bred for uniformity. So timing from birth to market weight (approximately 265 pounds) is reliably and reproducibly predicted. Ordinary mortality risks, primarily of piglets before or shortly following weaning is also well documented and predictable. Of course, mortality risks and the casualties caused by disease outbreaks alter these known patterns. The homogeneity of the hogs within an operation allows the progress of a cohort of pigs through the weight classes can be modeled as a function of time. One way to incorporate these would be to introduce into the model differential (estimable) survival rates that adjust the estimates according to the expected losses in each weight group.

Shocks, at least at affected operations, alter the equilibrium levels and distribution of hogs and pigs among the growth stages. Depending on the nature of the shock, hogs at different stages of maturity may be affected differently, depending on the virulence and the contagion of a disease, for example. As an illustration, reduced growths have been measured for hogs affected by pneumonia, which changes the transition rate from one weight class to the next. The production efficiency (or litter rate) and the number of hogs in lighter weight groups are impacted by infections such as leptospirosis, pseudorabies, PEVD and PRRS. More fatal diseases, such as Erysipelas and TGE, can significantly reduce the number of hogs in herds (if not wiping out all of them). On the other hand, the introduction of new vaccines, disease

containment procedures, feeding regimes, or genetic improvements may provide faster growth and accelerated transitions from one weight group to the next.

Especially in the early stages of a shock, the signal from the NASS survey is small. A potential shock may first be identified and/or the extent of operations affected not from the survey but through other sources, such as other USDA agencies or news of a major natural disaster. Thus, to be successful an approach must have the modeling flexibility to adjust the transition rates, as for example through incorporation of survival rates set by the joint use of expert-opinion and survey data.

2 KFM

2.1 KFM - Model

The first model (KFM) uses a time series approach with Kalman Filter to update the state of the system after each new observation is input (Durbin & Koopman 2012). This model combines information on hog inventory from multiple sources including survey measurements, inventory transaction data, relationship constraints, and Hog Board (ASB) analyst measurements. State-space representation is expressed through two system equations—a transition equation and an observation equation. These system equations describe the behavior of a condition or phase of a system.

Transition equations define how hog inventories change over time. Given the state of the system at a point in time, these equations determine the new state of the system at another/future point in time. Both linear and nonlinear equations are used to model the transitions for the hog model. Observation equations relate the state of the system with a set of measurements or observations from that state. Both linear and nonlinear modeling are utilized for the observation equations.

Five constraints are embedded in the model based on relationships between current and past inventory, the relationship of current and past inventory to external transactional data, and the hog growth cycle. These serve as an accounting system to ensure the consistency of entries both within and across quarters is implemented to track inventory increases (births, international imports) and inventory decreases (slaughter, death loss, international export).

1. **Death Loss** refers to the quantity of pig crop that dies after weaning and cannot be counted in the market weight groups. Therefore, Death Loss Ratio is the (annual) total pig crop divided by the (annual) total for weight groups 1 and 2 combined. This Ratio must exceed 1.0.

2. **Weight Group Transition** compares the (annual) total for weight group 1 plus a fixed fraction (α) of weight group 2 to the (annual) total for weight groups 3 and 4 plus the rest (1-α) of group 2. This ratio must exceed 1.0
3. **Pig Crop - Slaughter Ratio** constraint ensures the annual increase in slaughter is equal to the increase in births for the two preceding quarters. This results from the six-month time period between the birth and slaughter of a pig. This ratio is defined as the (two-quarter) total pig crop (minus death loss) divided by the total slaughter number. This ratio must equal 1.0.
4. **Market Hogs - Slaughter Ratio** constraint compares all market hogs (excepting breeding herd) with slaughter, with exception of breeding hogs. All market hogs (weight classes 1 through 4) will mature and go to slaughter within 6 months, so the total market hogs one quarter should equal the total slaughter numbers for the next two quarters combined.
5. **Market Hogs over 180lbs - Slaughter Ratio** constraint compares the number of hogs in weight class 4 (over 180lbs) to the slaughter number for the following quarter.

In addition, KFM fixes the survival rate (constant for all estimated weight classes).
Data used in the model includes the previous five quarters in conjunction with current data to capture annual cycle dynamics and an annual trend. Consequently the KFM model performs well in periods of stability or slow trend.

However the KFM model is unable to adapt quickly to sudden change due both to the built-in model stability (five previous quarters of data) and especially to the rigid constraints on the system. In consequence disturbance (e.g., disease outbreak) of previous relative equilibrium results in model estimates that are biased and may be quite unrealistic.
For a full and detailed description of the KFM model and the system of equations expressed in state-space form are see Busselburg (2013) in the Appendix.

2.2 KFM - Performance

The KFM model performs very well during times of equilibrium and is biologically sound. Constraints 4 and 5 ensure conformance to an external “gold standard” for national inventory numbers. Thus in terms of the performance criteria, as a predictive model the KFM meets criteria i) and iv) and partially satisfies iii).

However KFM, because of its relative inflexibility, fails to meet ii) and partially fails to meet iii). In the event of a shock, current input data does not override the rigid constraints plus the fixed survival rate. This results in a lag of at least a one quarter in detecting a shock. It should also be noted that the KFM is a national-scale model. Since shocks often are initially localized, the

effect of critical departures from system equilibrium can be damped down by data from the majority of states and extreme operations that are not as yet experiencing the disturbance.

Examining the performance of the KFM gives some insight into the relationships among the hog survey estimates, board numbers, and model estimates. In 2009-2013 there was a disease (PEDV) that adversely affected the hog population. This shock made inventory estimation difficult and often made estimates inaccurate. One consistent pattern during 2013 is evident by comparing the *final* board estimate with each of the contemporary estimates: All of the state recommendations, *initial* board, and KFM estimates overestimated the total hogs, i.e., they all underestimated the impact of the epidemic. As the epidemic waned, the contemporary estimates again lagged the event, in this case recovery. From September 2014 to June 2016 the state recommendations, *initial* board, and KFM estimates (with a few exception) were consistently below the *final* board estimate. Likely this resulted from inability to capture an upward trend as steep as is often seen with a population emerging from a downward shock. As the steep recovery largely resulted from operations greatly increasing the number of sows farrowed, diagnostics should be able to detect it. By 2016, hog production appears to have returned to equilibrium (Figure 1). From 2016 to the end of 2017 the KFM model estimates remain very close to *final* board estimate for the most part.

Figure 1 illustrates both the shortcomings and the strengths of the KFM model. During the period of shock (March 2013-March 2015) the average absolute deviation of the model estimate to the final board estimate was 1.48 million hogs, whereas during the steady state years (June 2015- December 2017) the average absolute deviation falls to .46 million hogs. This is additional evidence that the KFM model struggles to estimate total hogs during shocks. Figure 2 illustrates the absolute deviation from the *final* board estimate. Two patterns in this figure are important to note. First, the general downward slope of each line from 2013-2017 is caused by the high uncertainty during the shock years early on and afterward the system moving toward a steady state after 2015. This highlights the KFM's inability to provide accurate estimates during shocks and its ability to provide accurate estimates during steady state periods.

In summary the KFM was able to capture equilibrium picture accurately while satisfying external accounting relationships. However, it failed to adjust for shock as they appeared and was unable to account for disruptions in the birth to market hog lifecycle.

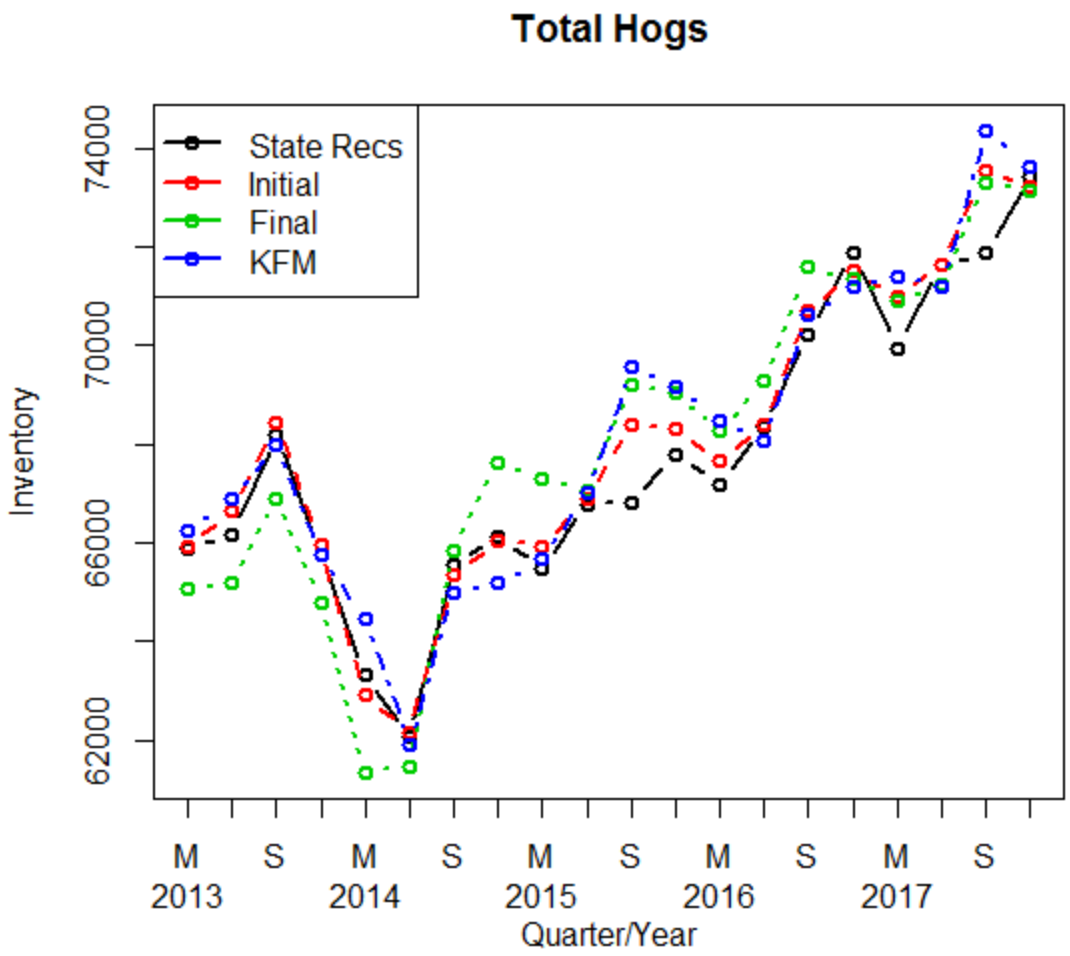

Figure 1 Illustrates the KFM estimate, state recs value, initial board estimate, and final board estimate for total hogs from March 2013 to December 2017.

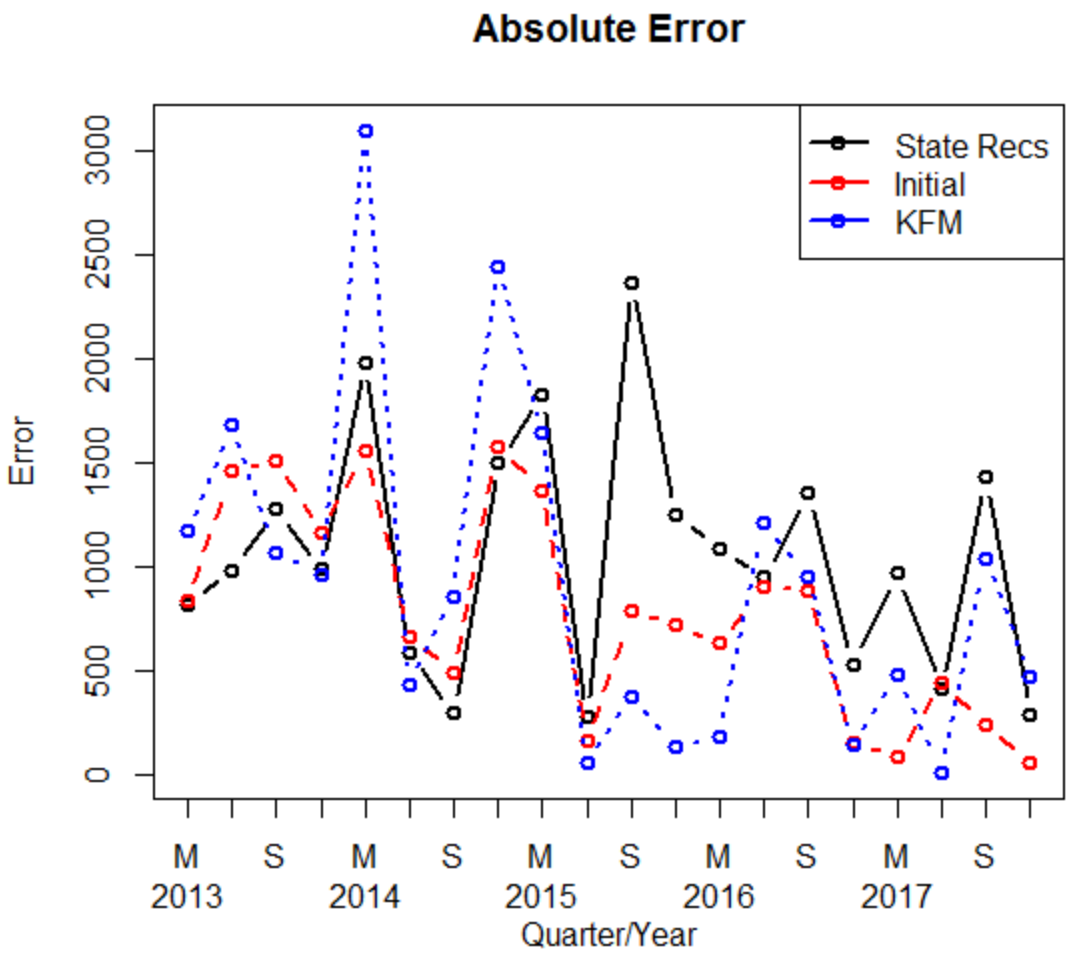

Figure 2 Illustrates the absolute error of state recs, initial board estimate, and the KFM with respect to the final board estimate

3 SGLM

3.1 SGLM – Model

Kedem and Pan (2015) developed a SGLM model for NASS in an attempt to address specifically the problem of periodic shocks that occur in hog inventories due to disease, natural disasters, tariffs, market forces, rapid structural changes, and new technologies. The choice of SGLM was based on the desire to give more weight to current data and immediate past in order to capture changing dynamics by giving more weight to the recent past. Another reason for adopting this "power house" modeling approach was to enable a dynamic covariate selection across a wide range of potential information including the survey and external data and adding economic information (such as hogs and pork prices). SGLM works by testing large numbers of potential

covariates using spectral analysis and then selecting among them for the final model according to their influence in explaining the output variables. Implementation was easy and fast via the web application *Shiny*. At the end of the process, estimates, forecasts and the measures of uncertainty are produced using the winning model.

3.2 SGLM- Performance

The SGLM has proved to be flexible, by its design; but without any constraints it is not stable even to the extent of incorporating a common core set of covariates from one quarter to the next. Thus it cannot satisfy either criterion iii) or iv) as there is no role in the SGLM model either for biologic relationships or for conformance to external administrative data.

The difficulties that SGLM encounters are acutely apparent during the immediate post-shock period, as illustrated in Figures 3 and 4 for the period from June 2016 through June 2017. The precision of SGLM model estimates could not consistently match the Hog Board's (ASB's) measure of accuracy, especially with the occurrence of a shock.

Furthermore, the SGLM had tremendous difficulty adjusting to the period immediately following the shock. This difficulty is illustrated in figures 3 & 4, between June 2015 and June 2016. The pattern of the error (compared to *final* board estimates) is a large initial increase during the shock, peaking immediately after but only beginning to decrease substantially one year after the shock.

At least as concerning as the increased error, is the inability of the model to adhere to the biological aspects of the hog life cycle. It has happened that the SGLM estimates for the number of hogs in the upper two weight classes is greater than the number of hogs in the lower weight classes the previous quarter (hogs that were set to transition into the upper weight classes). More commonly the flaw in the SGLM estimates was failure to account in future quarters for earlier losses in the lower weight classes. Table 1 provides an example from June 2016 where the total hogs from the SGLM does not equal the sum of its parts. Note that while the two bottom rows match for the KFM model, the discrepancy for the SGLM model is about 100,000 hogs. Also, the SGLM problem in this particular quarter has carry over effects into the following quarter because there is no relationship of hog inventory from weight group to weight group over time.

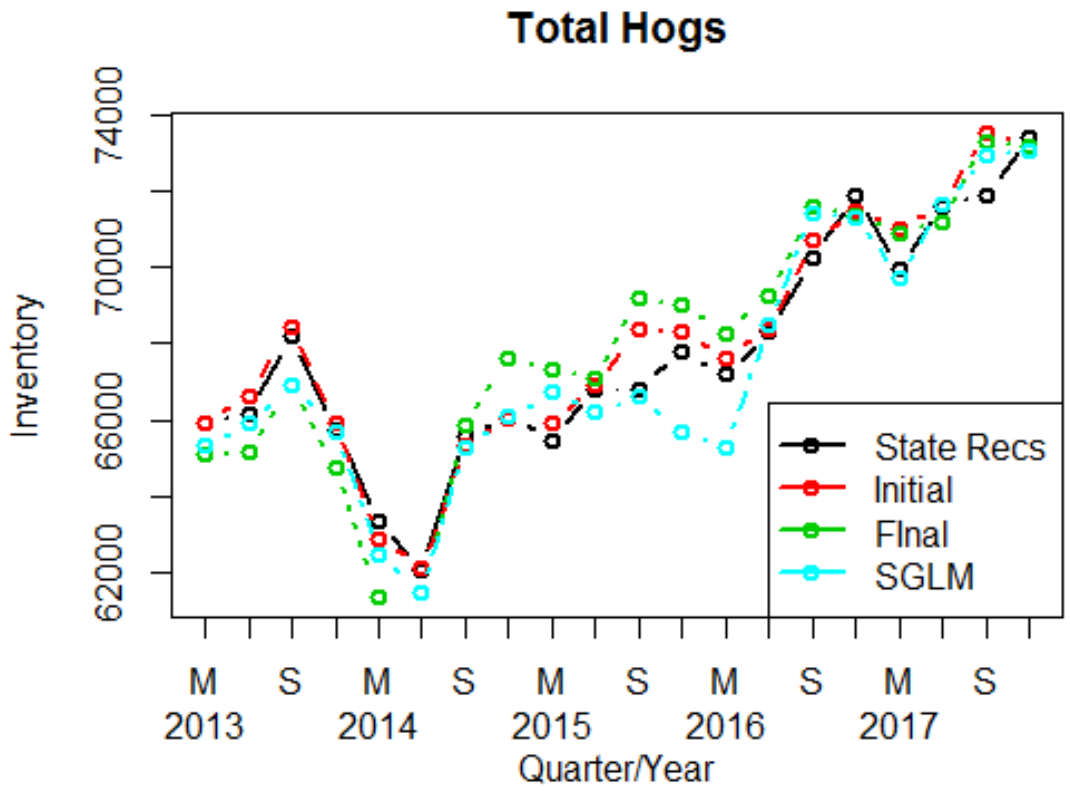

Figure 3 Illustrates the SGLM estimate, state recs value, initial board estimate, and final board estimate for total hogs from March 2013 to December 2017.

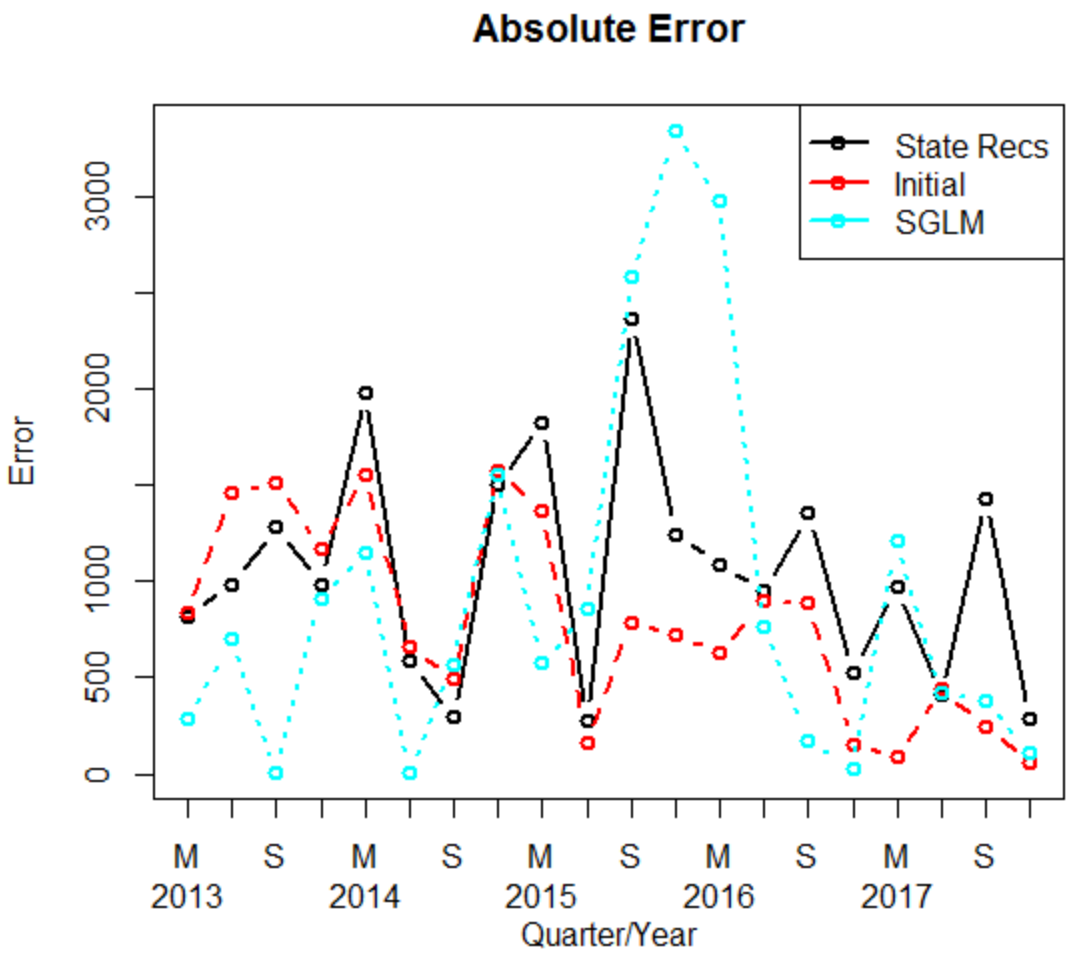

Figure 4 Illustrates the absolute error of state recs, initial board estimate, and the SGLM with respect to the final board estimate

Table 1. The relationship between the weight group and breeding herd estimates with the total hogs estimate from the KFM and the SGLM

		Model	
		SGLM	KFM
Estimate	G12	38943.00	39657.45
	G3	13324.15	13705.34
	G4	11258.00	11291.74
	BH	6011.72	5976.95
	Total Hogs	70656.00	70631.48
	Sum(G12,G3,G4,BH)	69536.87	70631.48

The SGLM model was only moderately successful, certainly not consistently successful, in producing total hog estimates during times of equilibrium and during shock years. However, the SGLM model is not a useful predictive mode, since the reliability of SGLM estimates require the input of the Pre-Board adjusted data.

4 Model Comparison

These two implemented models are capable of producing estimates with desirable characteristics but each has strengths the other lacks. While the KFM takes in consideration biological properties of hogs and captures an equilibrium dynamic that satisfies the accounting constraints, it is unable to adapt quickly to systematic shocks resulting in heavily biased and unrealistic results. On the other hand, SGLM model provides a flexible model that quickly captures the economic patterns and departures from an equilibrium state, but it does not satisfy reasonable biological dynamics of the hog population.

To improve these two approaches a flexible model is needed that takes into consideration the biological growth of hogs and tracks them from newborn piglets to market weight by modeling both their growth and their survival rates under different conditions (e.g. presence/absence of disease outbreaks). Other relationships, such as those between breeding herd and sows farrowed, require a separate formulation and when modeled over time provide indications of production changes.

Direct modeling of the biological properties of the system could allow elimination of the rigid biological constraints introduced by Busselberg (2013) while still producing reliable estimates and forecasts.

The biological aspect NASS wishes to incorporate into the modeling process would ideally mimic the basic life cycle and survival rates among hogs. This includes survival rates for hogs from birth into the initial weight groups, the survival among weigh groups, and the transition of hog cohorts across weight groups. The SGLM lacked the proper constraints to achieve these goals.

For now, the KFM model is the most useful tool currently available for use at NASS. Although the KFM model has some shortcomings, namely the inability to provide reliable estimates during shock periods (Figure 5), it has been reliable in periods emerging from shocks, when the SGLM model usually fails.

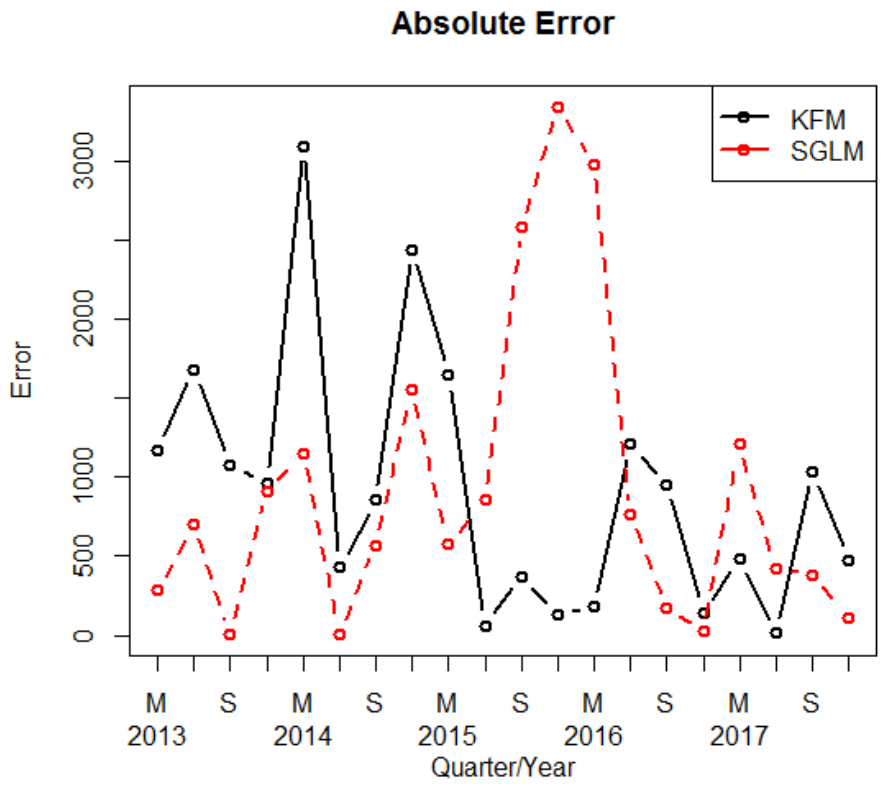

Figure 5 Absolute error of the KFM and SGLM models with respect to the final board estimate

5 Diagnostics

5.1 Shock Detection

In an effort to detect shocks as early as possible, Wang et.al. (2016) developed a Bayesian model to detect shocks for NASS in the hog and pig system as early as possible. Several variables were identified as useful in detecting the occurrence of a shock; these included initial (survey data) estimates of total hogs, sows farrowed and pig crops as well as differences in *initial, first* and *second* revisions of Hog Board estimates. (*Final* Board estimates were used to test the sensitivity of the diagnostics, using final Board estimates total hogs.)

Wang developed diagnostics for multiple-hypothesis testing of large scale (temporal) dependent data with the dependence structure among hypotheses being governed by a hidden Markov model (HMM). Their proposed testing procedure is based on Bayesian modeling framework, both parametric and non-parametric approaches.

These diagnostics were formulated by considering a Dirichlet mixture model with an unknown number of distributions for the non-null hypothesis. The state indicator probabilities then,

depending on the time of interest (i.e., after, at, or before the time of shock occurrence), can be described as the predictive state probability, filtered state probability, or smoothed state probability.

This algorithm allows for an optimal false negative rate, while controlling the false discovery rate (Wang et al., 2015). For full details, including the theory behind the applications see Wang et al. (2015) located in the appendix.

NASS uses the algorithm developed by Wang et al. (2015) to detect shocks occurring in the system. Each quarter, inventory estimates along with diagnostics like those shown in Figure 6 are provided for the livestock division. If there is a flag (shown as red dots) in the previous quarters for a possible shock then the Livestock Branch of the statistics Division (SD) is notified to be able to take this into consideration when setting official estimates.

For the known epidemic(s) with onset during 2013-2014, the diagnostics succeeded in flagging the quarters as shown from March 2013 through June 2014. (This is based on a reconstruction of the data and testing for those years prior to the actual development of the algorithm.) The limitation of these diagnostics is that, as currently employed, there is delay of one quarter (three months) in recognizing the earliest warning sign. For a predictive model it might be possible to compare the pre-data prediction with the initial model estimate (just as the initial estimate is compared now to the first revision) to eliminate the delay. For a predictive model it might be possible to compare the pre-data prediction with the initial model estimate (just as the initial estimate is compared now to the first revision) to eliminate the delay. Modification to the variance estimates would be required with the introduction of a purely predictive model based on previous quarters plus current information on farrowings and pig crop.

Data from December 2017 illustrate this point, as shown in Figure 6a. H_0 is the initial total hog estimate of December 2017; the first revision of the total hogs H_1 was produced in March 2018. Therefore this first date that has both H_0 and H_1 is September 2017. In other words, if a shock began in September 2017 diagnostics would not detect it until December 2017 when the first revision of total hogs for September 2017 was released. The example that follows uses estimates produced for the board on December 2017 (Figure 6).

5.2 Example

In the top panel of Figure 6, the diagnostic chart uses the data, H_1-H_0, where H indicates total hogs. These data are the Hog Board's *first* revision of total hogs for September 2017 minus the *initial* board estimate for total hogs of September 2017.

In the top panel, the last available data point is for September 2017; it is not red, so no shock is indicated. Similarly, the bottom panel of Figure 6 illustrates the diagnostic results using the data (H_3-H_0). This translates into using the Hog Board's *third* (revised) estimate of total hogs (as of March 2017) minus the *initial* estimate of total hogs (September- a lag of three quarters. The earliest available indication of shock based on these data is March 2017 and there is no indication of shock during this time. The higher number of flags (red dots) on the bottom panel compared to the top panel most likely relates to the Hog Board's corrections over the period of shock. As a general observation, during shocks each subsequent revision tends become more distant from the *initial* estimate presumably because more information continues to emerge about the extent of the shock.

The diagnostics are the final piece of information provided to the board for them to make informed estimates of hog inventory. Diagnostic results, together with survey estimates, slaughter data, historical data, model estimates, and state recommendations are evaluated and balanced to produce biologically sound board estimates. No single number or indication is used to produce final board estimates; it is a carefully balanced and controlled process that involves an assembly of information.

For this example, the Research and Development Division's deliverables to Statistics Division are the model estimates and statement that as of September 2017 there is no indication of shocks.

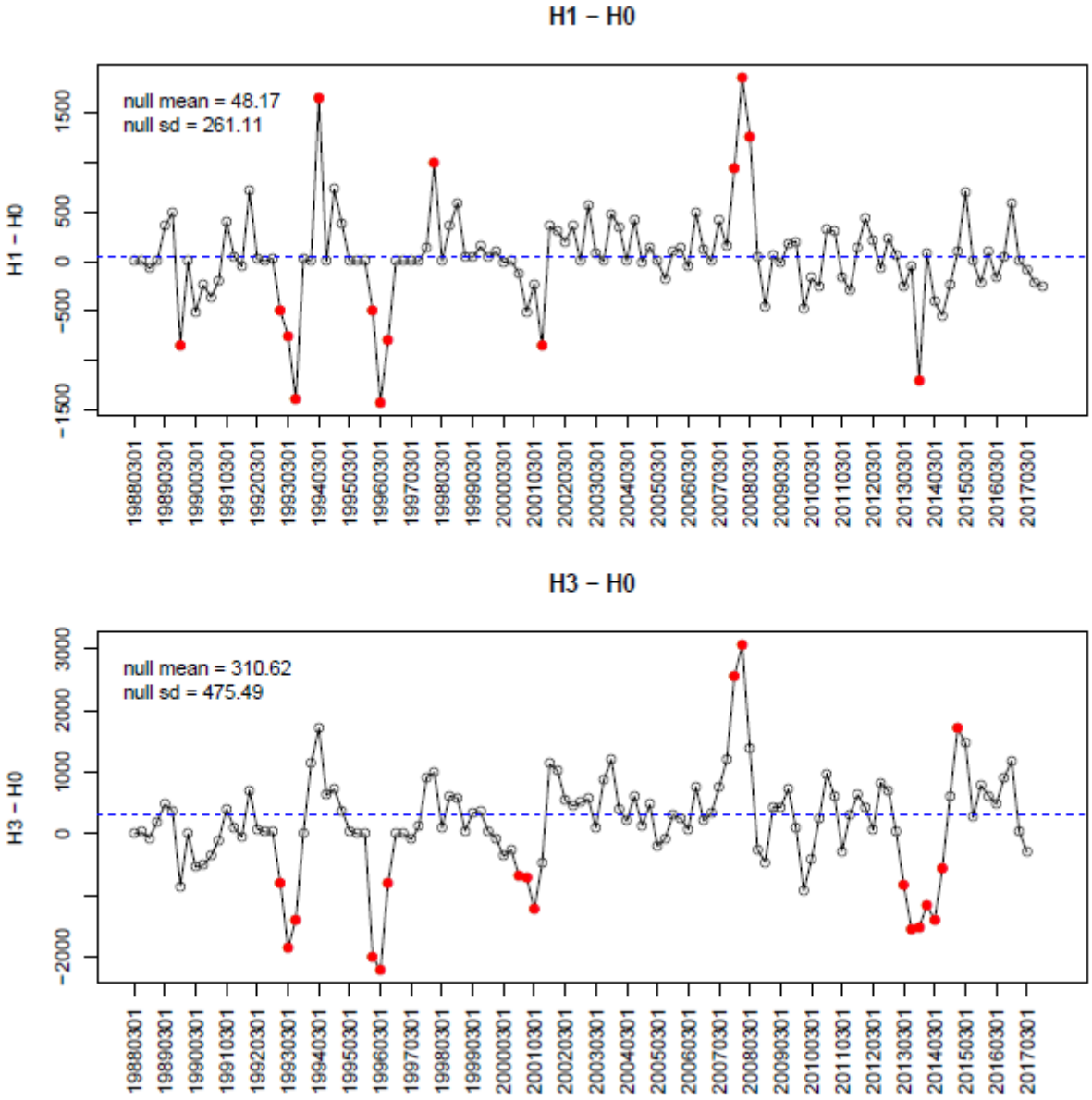

Figure 6 Illustrates the output for detecting shocks. This output uses revised total hog numbers as inputs. Time periods marked in red "flag" potential occurrences of shocks.

6 Work Cited

Busselberg, S. (2013) The Use of Signal Filtering for Hog Inventory Estimation. United States Department of Agriculture NASS.

Durbin, J. and Koopman, S.J. Time Series Analysis by State Space Methods. Oxford: Oxford University Press, 2012.

Drabenstott, M. (1998). This Little Piggy Went to Market: Will the New Pork Industry Call the Heartland Home?

Fèvre, E. M., Bronsvoort, B. M. D. C., Hamilton, K. A., & Cleaveland, S. (2006). Animal movements and the spread of infectious diseases. *Trends in microbiology*, *14*(3), 125-131.

Hubbell, B. J., & Welsh, R. (1998). An examination of trends in geographic concentration in US hog production, 1974–96. *Journal of Agricultural and Applied Economics*, *30*(2), 285-299.\

Kedem, B & Pan, L. (2015) *Time Series Prediction of Hog Inventory.* Unpublished internal document. United States Department of Agriculture NASS.

Rhodes, V. J. (1995). The industrialization of hog production. *Review of Agricultural economics*, 107-118.

Sullivan, J., Utpal, V., & Smith, M. (2000). Environmental regulation & location of hog production. *Agricultural Outlook*, (274), 19-23.

Valdes-Donoso, P., VanderWaal, K., Jarvis, L. S., Wayne, S. R., & Perez, A. M. (2017). Using Machine learning to Predict swine Movements within a regional Program to improve control of infectious Diseases in the US. *Frontiers in veterinary science*, *4*, 2.

Chapter 5: Modeling swine population dynamics

Luca Sartore

1 Overview

The proposed model provides monthly estimates of hog inventories at a national level by modeling biologic dynamics for the US swine population. The main variables used by the proposed methodology are briefly described in Table 1. The estimates $\hat{\mathbf{y}}_t$ are obtained at the end of a sequence of five processes:

- **Initial information**: gathering preliminary information in numerical format;
- **Pre-processing**: adjusting and summarizing the initially gathered information into a single dataset;
- **First estimation**: producing estimates for the pre-board;
- **Information update**: updating the dataset to be used in the next step;
- **Second estimation**: producing updated estimates for the Agriculture Statistics Board (ASB).

These five processes are depicted in Figure 1.

The first stage (*initial information*) consist of gathering initial information that will be organized and made available for computations. This initial information consists of:

- the micro-data $\mathbf{y}_t$ from survey respondents at time t,
- the NASS published estimates based on historical information, and
- the state recommendations provided by NASS field offices.

Next, the *pre-processing* stage fundamentally consists of three main operations that are performed with the purpose of generating a comprehensive dataset that accounts for both the historical dynamics of the hog population and the survey data. During this stage, micro-data are aggregated into summary statistics that are adjusted to reduce bias and improve the final results.

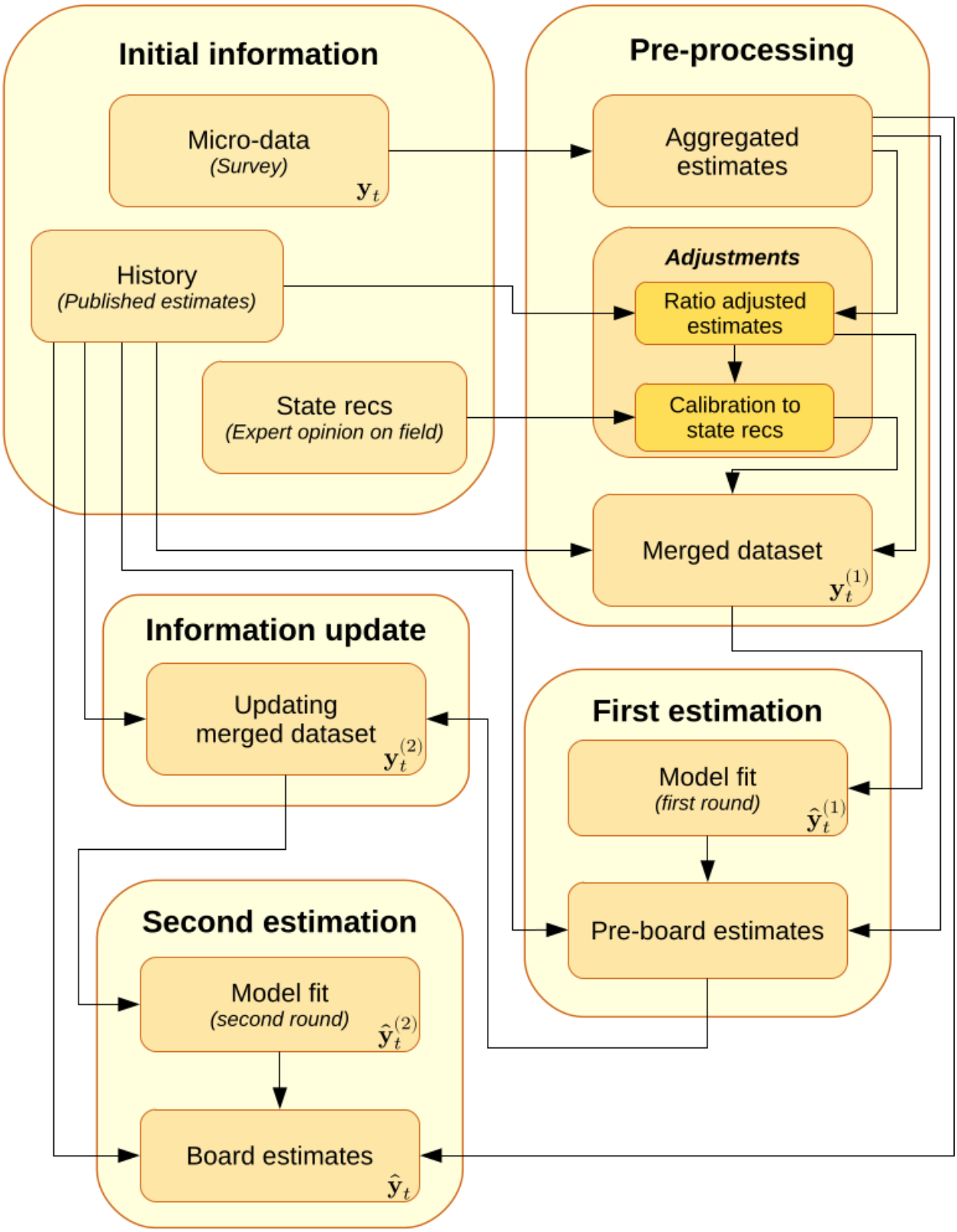

Figure 1: Estimation process.

Once a comprehensive dataset $\mathbf{y}_t^{(1)}$ is created, the *first estimation* process starts. The parameters of the new model are estimated by using iterative regression techniques, and the fitted values for the variables of interest $\hat{\mathbf{y}}_t^{(1)}$ are calculated for the most recent quarter. The output from the estimation procedure is then passed to the pre-board along with historical data and aggregated survey data. Four experts forming the pre-board assess the available information and set estimates that account for relevant factors not captured by the modeled dynamics and/or the survey.

The pre-board provides a set of estimates $\mathbf{y}_t^{(2)}$ for the current quarter and revised values of the official statistics that are used to update the values produced at the pre-processing stage (*information update*) and those provided by the historical records to be consistent with other administrative sources of information.

The dataset used in modeling is revised to reflect the pre-board estimates. Then the *second estimation* process begins. This final procedure consists of two consecutive steps. First, the model is fitted by using the updated dataset as input. Second, the results from the model $\hat{\mathbf{y}}_t^{(2)}$ are provided to the ASB, consisting of nine or ten livestock-commodity experts (including those forming the pre-board) who set the official estimates $\hat{\mathbf{y}}_t$ for the current quarter.

Further details on the full estimation process are provided in the next sections.

Table 1: Notation of the main variables used by the proposed methodology.

Variable description	Survey data	Adjusted estimates	First model estimates	Pre-board estimates	Second model estimates	Final board estimates
Pig crop (piglets)	$y_{1,t}$	$y_{1,t}^{(1)}$	$\hat{y}_{1,t}^{(1)}$	$y_{1,t}^{(2)}$	$\hat{y}_{1,t}^{(2)}$	$\hat{y}_{1,t}$
Sows farrowed	$y_{2,t}$	$y_{2,t}^{(1)}$	$\hat{y}_{2,t}^{(1)}$	$y_{2,t}^{(2)}$	$\hat{y}_{2,t}^{(2)}$	$\hat{y}_{2,t}$
Breeding herd	$y_{3,t}$	$y_{3,t}^{(1)}$	$\hat{y}_{3,t}^{(1)}$	$y_{3,t}^{(2)}$	$\hat{y}_{3,t}^{(2)}$	$\hat{y}_{3,t}$
Weight group 1	$y_{4,t}$	$y_{4,t}^{(1)}$	$\hat{y}_{4,t}^{(1)}$	$y_{4,t}^{(2)}$	$\hat{y}_{4,t}^{(2)}$	$\hat{y}_{4,t}$
Weight group 2	$y_{5,t}$	$y_{5,t}^{(1)}$	$\hat{y}_{5,t}^{(1)}$	$y_{5,t}^{(2)}$	$\hat{y}_{5,t}^{(2)}$	$\hat{y}_{5,t}$
Weight group 3	$y_{6,t}$	$y_{6,t}^{(1)}$	$\hat{y}_{6,t}^{(1)}$	$y_{6,t}^{(2)}$	$\hat{y}_{6,t}^{(2)}$	$\hat{y}_{6,t}$
Weight group 4	$y_{7,t}$	$y_{7,t}^{(1)}$	$\hat{y}_{7,t}^{(1)}$	$y_{7,t}^{(2)}$	$\hat{y}_{7,t}^{(2)}$	$\hat{y}_{7,t}$
Vector including all variables	$\mathbf{y}_t$	$\mathbf{y}_t^{(1)}$	$\hat{\mathbf{y}}_t^{(1)}$	$\mathbf{y}_t^{(2)}$	$\hat{\mathbf{y}}_t^{(2)}$	$\hat{\mathbf{y}}_t$

2 Data adjustments

2.1 Aggregation

Table 2: Notation used in Section 2.1.

Notation	Description
t	Index denoting time on a monthly basis.
i	Index denoting the i-th hog operation.
k	Index denoting the variable considered (see below for details).
$k = 1$	Index denoting pig crop.
$k = 2$	Index denoting sows farrowed.
$k = 3$	Index denoting weight group 1 (hogs below 50 lbs).
$k = 4$	Index denoting weight group 2 (hogs between 50 lbs and 119 lbs).
$k = 5$	Index denoting weight group 3 (hogs between 120 lbs and 179 lbs).
$k = 6$	Index denoting weight group 4 (hogs above 180 lbs).
$k = 7$	Index denoting breeding herd.
$y_{k,t,i}$	The value of variable k at time t for the operator i.
$w_{t,i}$	Weight associated with the record i for the survey conducted at time t.
n_t	Number of hog operations sampled at time t.
$N_{j,t}$	Number of hog operations in the list frame of stratum j at time t.
$n_{j,t}$	Number of hog operations sampled from (the list frame) stratum j at time t.
$a_{j,t}$	Number of respondent operations within the sample of stratum j at time t.
$\pi_{t,i}^{(S)}$	Sample inclusion probability associated with record i at time t.
$\pi_{t,i}^{(C)}$	Coverage probability associated with record i at time t.
$\pi_{t,i}^{(R)}$	Response probability associated with record i at time t.

Micro data $y_{k,t,i}$ are aggregated for each variable $k = 1, \dots, 7$, by computing the weighted sum with the survey weights $w_{t,i}$ accounting for the incompleteness of the list frame, the probability that unit i at time t is included in the sample, and the lack of response from some sampled units (NASS 2005). The design-based estimate of the national total of variable k is then

$$\sum_{i=1}^{n_t} w_{t,i}\, y_{k,t,i}. \qquad (1)$$

The sampling weight for each record in stratum j at time t is the proportion of the units in stratum j that are included in the sample:

$$\pi_{t,i}^{(S)} = \frac{n_{j,t}}{N_{j,t}}. \qquad (2)$$

As described in Chapter 3, the June Area Survey (JAS) records that are not on the NASS list frame (NOL) are used to assess under-coverage. The NOL records are included in the December sample, and the weighted proportion of records on the list frame within stratum j at time t is the coverage probability for that stratum at that time. In March, June, and September the probability of coverage is modeled.

The probability of response for records in stratum j at time t is estimated by the proportion of responding records in that stratum at time t:

$$\pi_{t,i}^{(R)} = \frac{a_{j,t}}{n_{j,t}}. \qquad (3)$$

The survey weight of record i at time t, i.e. $w_{t,i}$, is then the reciprocal of the product of the probability of being in the sample and the estimated probabilities of coverage and response. The survey estimates tend to be biased downward (see Chapter 3). Before using them in the modeling process, the design-based estimates are adjusted so that they are consistent with the state recommended estimates (see Chapter 2). The adjustment is a two-step process. First, to adjust the bias in the survey estimates, the historical relationships between the official estimates and the corresponding survey estimates are used to construct ratio-adjusted estimates for the current quarter. Because state-recommended estimates are a major factor in setting the preliminary board estimates, the ratio-adjusted estimates are calibrated to the state recommended estimates. More details and motivation about the ratio adjustments are provided in the following sections.

2.2 Ratio adjustments

Table 3: Notation used in Section 2.2.

Notation	Definition
$y_{k,t}$	Survey estimate for variable k at time t.
$y_{k,t}^{(1)}$	Adjusted estimate for variable k at time t.
$\widehat{y}_{k,t}$	Official board estimate for variable k at time t.
$\hat{z}_{1,t}$	ASB monthly estimate of pig crop at time t.
$\hat{z}_{2,t}$	ASB monthly estimate of sow farrowed at time t.
$r_{k,t}$	Estimated ratio for adjusting the variable k at time t.
t_0	Current estimation time (month correspondent to the quarter when survey is conducted).
$H_{s,u}$	Value associated with neural-network neuron u at layer s.
$\eta_{s,u,0}$	Intercept parameter associated with neuron u at layer s.
U_{s-1}	Number of neurons at the previous layer.
$\eta_{s,u,v}$	Parameter associated with output neuron u at layer s weighting input neuron v at layer $s-1$.
$H_{s-1,v}$	Value associated with neuron v at the previous layer.

As discussed in Chapter 3, the survey-based estimates are biased downward and thus consistently below the official estimates. The estimation of the ratios of the current variables of interest and the corresponding survey estimates is based on the observed ratios between the past official estimates $\widehat{y}_{k,t}$, and the corresponding survey estimates $y_{k,t}$. The current questionnaire was introduced in March 2008 so only data from that time to the present are used. The ratios constructed from historical records are modeled to estimate the ratios for the current time $t = t_0$, i.e. r_{k,t_0}, and subsequently the ratio-adjusted estimates $y_{k,t_0}^{(1)} = r_{k,t_0}\ \widehat{y}_{k,t_0}$.

Recall that monthly data are reported for pig crop and sows farrowed ($k = 1$ and 2), and quarterly data are reported for the other variables ($k = 3, \ldots, 7$). First, consider the past monthly observed ratios for pig crop $\hat{z}_{1,t}$ and sows farrowed $\hat{z}_{2,t}$, where time t is expressed on a monthly basis. Quarterly estimates of pig crop and sow farrowed produced at time t, respectively, refer to the total numbers of weaned piglets and farrowing sows reported during the months $t-1$, $t-2$, and $t-3$. Thus, the quarterly official estimates of the variables $k = 1,2$ are computed as

$$\widehat{y}_{k,t} = \sum_{h=1}^{3} \hat{z}_{k,t-h}\,, \qquad (4)$$

e.g. the quarterly value of $y_{k,t}$ in September is the sum of the reported values on a monthly basis: $z_{k,t-1}$ in August, $z_{k,t-2}$ in July, and $z_{k,t-3}$ in June, for $k = 1$, and 2, respectively.

The estimated ratio adjustments for the variable $k = 1,2$, for $t-1$, $t-2$, and $t-3$, are computed by using all the information available, i.e. $r_{k,t-h} = \widehat{y}_{k,t-h} / y_{k,t-h}$, where $h = 1,2, \ldots$ presents months prior to time $t = t_0 - 3$. These values are then used to fit a model that predicts the estimated ratio adjustments for variables $k = 1$, and 2, at time $t = t_0 - 1$, $t_0 - 2$, and $t_0 - 3$.

For each variable $k = 3, \ldots, 7$, the estimated ratio adjustments are computed over the quarters, i.e. $r_{k,t-3\ell} = \widehat{y}_{k,t-3\ell} / y_{k,t-3\ell}$, where $\ell = 1,2, \ldots$ counts the previous quarters from time $t = t_0$. For example, for the March quarterly report, the weaned piglets and sows farrowed in December of the preceding year and January and February of the current year are of interest. Notice that this notation has been developed to have a direct connection between quantities expressed on a monthly basis with those expressed on a quarterly basis. One can develop (and evaluate) a model capable of predicting the estimated ratios r_{k,t_0} for the current quarter.

The proposed model adopted for this task is based on a single neural network with two hidden layers. This model extracts quarter-to-quarter nonlinear feature dynamics of the historical ratios. The two hidden layers consists of three and two neurons (see for example Figure 2).

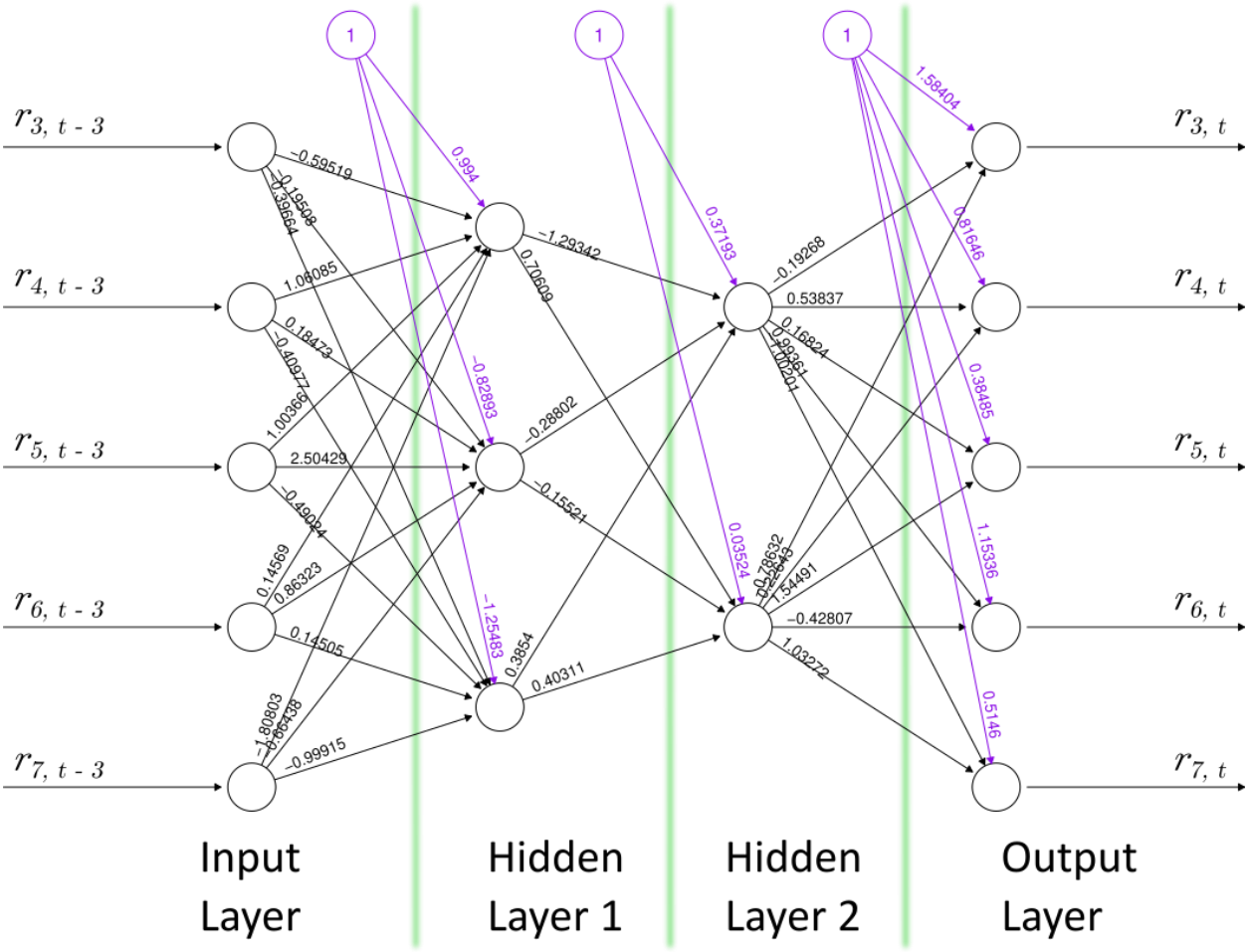

Figure 2: Example of a neural network with two hidden layers of three and two neurons each.

As shown in Figure 2, the information (i.e. the ratio adjustments for several variables) of a past quarter $t-3$ flows from the input layer through the hidden layers, and it is linearly combined into the output layer, which produces the predicted ratio for quarter t. This flow is formulated as

$$H_{s,u} = \left[1 + \exp\left(-\eta_{s,u,0} - \sum_{v=1}^{U_{s-1}} \eta_{s,u,v}\, H_{s-1,v}\right)\right]^{-1}, \qquad (5)$$

for layer $s = 1,2$. In particular, the values associated with the input-layer neurons $H_{0,v}$ are equivalent to the estimated ratio adjustments at the previous quarter, i.e. $r_{k,t-3}$ in Figure 2. The values associated with the output-layer neurons

$$H_{3,u} = \eta_{3,u,0} + \eta_{3,u,1} H_{2,1} + \eta_{3,u,2} H_{2,2} \qquad (6)$$

are the estimated ratio adjustments at time t. Typically, the parameters $\eta_{s,u,v}$ are estimated by back-propagation algorithms introduced by Rumelhart, Hinton, and Williams (1985).
Neural network models have been chosen for their ability to extract intricate nonlinear features, reduce highly dimensional spaces, and expand smaller ones. These models have been successfully applied in many fields by providing accurate predictions for describing non-linear phenomena. They can be very flexible and over-parameterized due to their recursive nature as defined in (5). However, using two layers with a small number of neurons can provide reliable results with minimal computational effort.

2.3 Calibration to state recommendations

Table 4: Notation used in Section 2.3.

Notation	Definition
$y_{k,t}^{(1)}$	Adjusted estimate for the variable k at time t.
$y_{\bullet,t}^{*}$	State recommendation for total hogs at time t.
$r_{k,t}$	Ratio adjustment for variable k at time t (output of the neural network).
$\hat{r}_{k,t}$	Calibrated value of the ratio adjustment for variable k at time t.
t_0	Current estimation time.
λ	Lagrange multiplier.

The NASS field offices provide recommended estimates for the total number of hogs in each state. These numbers are then summed to a national total, which is equivalent to $y_{\bullet,t_0}^{*}$. The estimates provided by the field offices incorporate administrative data collected by State Departments of Agriculture. The state-recommended estimates are used to further adjust the ratio-adjusted estimates $y_{k,t_0}^{(1)}$, for $k = 3, \ldots, 7$, through calibration. The aim of the calibration process is to match the adjusted sample estimates of national total inventory to the total based on the state recommendations by minimizing the following quantity:

$$\sum_{k=3}^{7} \left(\hat{r}_{k,t_0} - r_{k,t_0}\right)^2, \qquad (7)$$

which denotes the Euclidean distance between the calibrated ratio adjustments $\hat{r}_{k,t_0}$ and the estimated ratio adjustments r_{k,t_0} obtained through the neural network.

Calibration alters the ratio adjustments produced by the neural network to satisfy the following constraint:

$$y^*_{\bullet,t_0} = \sum_{k=3}^{7} \hat{r}_{k,t_0}\, y_{k,t_0}. \qquad (8)$$

The state-recommended estimate of US total number of hogs produced by the state recommendations on the left side of equation (8) serves as a benchmark for the calibrated sample estimate of the total national inventory, i.e. the right side of equation (8), where

$$\hat{r}_{k,t_0}\, y_{k,t_0} = y^{(1)}_{k,t_0}, \qquad (9)$$

for $k = 3, \ldots, 7$. Litter rate and sows farrowed are not calibrated because there are no state recommendations for these.

The calibration problem can be solved by finding the stationary point of the following quantity with respect to each $\hat{r}_{k,t_0}$ and λ:

$$\sum_{k=3}^{7} \left(\hat{r}_{k,t_0} - r_{k,t_0}\right)^2 + \lambda \left(\sum_{k=3}^{7} \hat{r}_{k,t_0}\, y_{k,t_0} - y^*_{\bullet,t_0} \right), \qquad (10)$$

where λ is a scalar used as a penalty if the equality constraint set by equation (8) does not hold. By taking the derivatives of (10) with respect to each $\hat{r}_{k,t_0}$ and λ and setting them to zero, one obtains the following system of equations to solve:

$$\begin{cases} 2\left(\hat{r}_{k,t_0} - r_{k,t_0}\right) + \lambda y_{k,t_0} = 0, \quad \forall k = 3, \ldots, 7, \\ \sum_{k=3}^{7} \hat{r}_{k,t_0}\, y_{k,t_0} - y^*_{\bullet,t_0} = 0. \end{cases} \qquad (11)$$

Thus, the optimal ratio adjustments can be estimated through the following equation:

$$\hat{r}_{k,t_0} = r_{k,t_0} - \frac{\lambda y_{k,t_0}}{2}, \qquad (12)$$

for $k = 3, \ldots, 7$, where the Lagrange multiplier λ is computed by substituting equation (12) into the last equation of the system (11). By solving for λ, the resulting formula for the Lagrange multiplier can be expressed as

$$\lambda = 2\,\frac{\sum_{k=3}^{7} r_{k,t_0}\, y_{k,t_0} - y^*_{\bullet,t_0}}{\sum_{k=3}^{7} y^2_{k,t_0}}. \qquad (13)$$

By combining the results of equations (12) and (13), the estimated ratio adjustments for variables $k = 3, \ldots, 7$ are computed directly by the use of the following closed-form solution:

$$\hat{r}_{k,t_0} = r_{k,t_0} - y_{k,t_0}\,\frac{\sum_{k=3}^{7} r_{k,t_0}\, y_{k,t_0} - y^*_{\bullet,t_0}}{\sum_{k=3}^{7} y^2_{k,t_0}}. \qquad (14)$$

3 The new model

The SWARCS model (named after the initials of the authors' surnames) tracks the growth of the surviving newborn piglets in the population and provides monthly estimates for the inventory number of market hogs (classified by weight). The estimates for the size of the breeding herd, the pig crop (i.e. number of weaned piglets), and the number of sows farrowed are also provided. The model is based on the assumption of an average dynamic growth rate for weaned pigs born within a month, and considers standard practices of the swine industry.

A conceptual map of the hog production chain can be used to formulate class transitions and relationships among quantities to be modeled. The evolution of the hog production system can be visualized by considering classical approaches used by managers to establish and improve the efficiency of processes. This analysis leads to a simple model that characterizes the connections among the variables of interest (see Figure 3). The estimates honor biological constraints.

The model is divided into two system of equations:

- The first describes the relationships between sows farrowed and pig crop, which are measured on a monthly basis. The number of sows farrowed is also related to the size of the breeding herd for the previous quarter. These numbers are available through the quarterly surveys on a monthly basis and can be used to track hog production on a finer time resolution to provide quarterly inventory estimates.
- The second defines the total inventories of four weight groups at the national level. These totals, together with the size of the breeding herd, form the total hogs in the US. The four weight groups are accounted for in this second system of equations; however, the size of

the breeding herd is part of the first system due to the close relationship with the number of sows farrowed (see Figure 3).

Although all data are counts of the number of heads, a normal approximation to these counts is used due to:

- the large numbers of hogs and pigs to estimate, and
- the computational simplification of the estimating equations.

This approach will be considered and explained in the following sections.

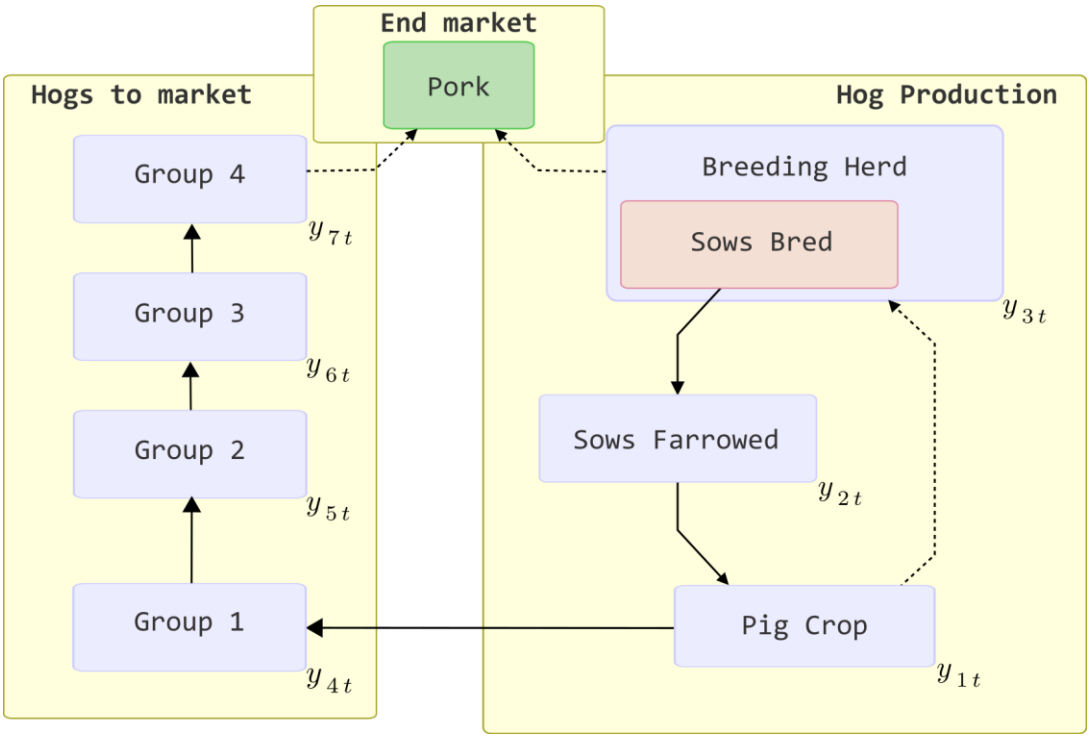

Figure 3: Pork production processes.

3.1 Model for monthly estimates

Table 5: Notation used in Section 3.1.

Notation	Description
$z_{1,t}$	Monthly pig-crop at time t.
$z_{2,t}$	Monthly sows farrowed at time t.
$y_{7,t-2}$	Breading herd at time $t-2$.
ρ_t	Litter rate at time t
φ	Farrowing rate.
$\varepsilon_{2,t}$	Statistical error in modeling sows farrowed.
B^h	Backward operator of h steps. E.g. the notation $B^3 z_{1,t}$ is equivalent to $z_{1,t-3}$.
∇_S^d	Difference operator of order d at lag S. E.g. the notation $\nabla_{12}^2 z_{1,t}$ is equivalent to $(1-B^{12})^2 z_{1,t} = z_{1,t} - 2z_{1,t-12} + z_{1,t-24}$.
$\tilde{\varepsilon}_{1,t}$	Statistical error in modeling the logarithm of pig crop $\log(z_{1,t})$.
$\tilde{\varepsilon}_{2,t}$	Statistical error in modeling the logarithm of sows farrowed $\log(z_{2,t})$.
$\tilde{\varepsilon}_{\rho,t}$	Statistical error in modeling the logarithm of litter rate $\log(\rho_t)$.
ϕ_i	Autoregressive parameter associated with time-lag i.
θ_i	Moving-average parameter associated with time-lag i.

At the national level, pig crop, sows farrowed and litter rate are modeled differently than the five basic inventory items:

- Market hogs weighting less than 50 lbs.
- Market hogs weighting between 50 and 119 lbs.
- Market hogs weighting between 120 and 179 lbs.
- Market hogs weighting 180 lbs or more.
- Breeding herd, including sows kept for breeding.

The strategy of having two separate models allows NASS to account for different time units (monthly for pig crop and sows farrowed versus quarterly for the others) and provides a reasonable explanation of the hog population dynamics from a macroscopic perspective.

The equations governing the number of sows farrowed and the pig crop are the following:

$$\begin{cases} z_{1,t} = \rho_t \, z_{2,t}, \\ z_{2,t} = \varphi \, y_{7,t-2} + \varepsilon_{2,t}, \end{cases} \qquad (15)$$

where the logarithm of pig crop $\log(z_{1,t})$ and the logarithm of sows farrowed $\log(z_{2,t})$ are each modeled by a Seasonal AutoRegressive Integrated Moving Average (SARIMA) model (Box et al. 2015). In particular, a SARIMA$(2,1,2) \times (2,1,2)_{12}$ is fit using LASSO regression (Tibshirani 1996). The LASSO shrinks the parameter estimates for some variables toward zero by the use of a penalty term that is added to the likelihood. The variables with parameter estimates of zero are removed, resulting in a parsimonious model with the remaining variables being most closely associated with the response. This approach, as shown by Wang, Li, and Tsai (2007), also allows for automatic time series model selection to be used in the estimation of the logarithms of pig crop, $\log(z_{1,t})$, and sows farrowed, $\log(z_{2,t})$. Thus, in addition to equation (15), the following set of equations should be considered in the estimation process:

$$\begin{cases} \log(z_{1,t}) = \dfrac{\left(1+\theta_{1,1}B+\theta_{1,2}B^2\right)\left(1+\theta_{1,12}B^{12}+\theta_{1,24}B^{24}\right)\tilde{\varepsilon}_{1,t}}{\left(1+\phi_{1,1}B+\phi_{1,2}B^2\right)\left(1+\phi_{1,12}B^{12}+\phi_{1,24}B^{24}\right)\nabla_1\nabla_{12}}, \\ \log(z_{2,t}) = \dfrac{\left(1+\theta_{2,1}B+\theta_{2,2}B^2\right)\left(1+\theta_{2,12}B^{12}+\theta_{2,24}B^{24}\right)\tilde{\varepsilon}_{2,t}}{\left(1+\phi_{2,1}B+\phi_{2,2}B^2\right)\left(1+\phi_{2,12}B^{12}+\phi_{2,24}B^{24}\right)\nabla_1\nabla_{12}}, \\ \log(\rho_t) = \dfrac{\left(1+\theta_{1,1}B+\theta_{1,2}B^2\right)\left(1+\theta_{1,12}B^{12}+\theta_{1,24}B^{24}\right)\tilde{\varepsilon}_{1,t}}{\left(1+\phi_{1,1}B+\phi_{1,2}B^2\right)\left(1+\phi_{1,12}B^{12}+\phi_{1,24}B^{24}\right)\nabla_1\nabla_{12}} - \\ \quad - \dfrac{\left(1+\theta_{2,1}B+\theta_{2,2}B^2\right)\left(1+\theta_{2,12}B^{12}+\theta_{2,24}B^{24}\right)\tilde{\varepsilon}_{2,t}}{\left(1+\phi_{2,1}B+\phi_{2,2}B^2\right)\left(1+\phi_{2,12}B^{12}+\phi_{2,24}B^{24}\right)\nabla_1\nabla_{12}}. \end{cases} \tag{16}$$

Therefore, the expanded form for the logarithm of pig crop at time t is

$$\begin{aligned} \log(z_{1,t}) = {} & -\phi_{1,2}\phi_{1,24}\log(z_{1,t-39}) + \left(\phi_{1,2}-\phi_{1,1}\right)\phi_{1,24}\log(z_{1,t-38}) + \left(\phi_{1,1}-1\right)\phi_{1,24}\log(z_{1,t-37}) + \\ & + \left(\phi_{1,2}\phi_{1,24}-\phi_{1,12}\phi_{1,2}\right)\log(z_{1,t-27}) + \left(\left(\phi_{1,1}-\phi_{1,2}\right)\phi_{1,24}+\phi_{1,12}\phi_{1,2}-\phi_{1,1}\phi_{1,12}\right)\log(z_{1,t-26}) + \\ & + \left(\left(1-\phi_{1,1}\right)\phi_{1,24}+\left(\phi_{1,1}-1\right)\phi_{1,12}\right)\log(z_{1,t-25}) + \left(\phi_{1,12}-\phi_{1,24}\right)\log(z_{1,t-24}) + \\ & + \left(\phi_{1,12}-1\right)\phi_{1,2}\log(z_{1,t-15}) + \left(\left(1-\phi_{1,12}\right)\phi_{1,2}+\phi_{1,1}\phi_{1,12}-\phi_{1,1}\right)\log(z_{1,t-14}) + \\ & + \phi_{1,24}\log(z_{1,t-36}) + \left(\left(1-\phi_{1,1}\right)\phi_{1,12}+\phi_{1,1}-1\right)\log(z_{1,t-13}) + \left(1-\phi_{1,12}\right)\log(z_{1,t-12}) + \\ & + \phi_{1,2}\log(z_{1,t-3}) + \left(\phi_{1,1}-\phi_{1,2}\right)\log(z_{1,t-2}) + \left(1-\phi_{1,1}\right)\log(z_{1,t-1}) + \theta_{1,2}\theta_{1,24}\,\tilde{\varepsilon}_{1,t-26} + \\ & + \theta_{1,1}\theta_{1,24}\,\tilde{\varepsilon}_{1,t-25} + \theta_{1,24}\,\tilde{\varepsilon}_{1,t-24} + \theta_{1,12}\theta_{1,2}\,\tilde{\varepsilon}_{1,t-14} + \theta_{1,1}\theta_{1,12}\,\tilde{\varepsilon}_{1,t-13} + \theta_{1,12}\,\tilde{\varepsilon}_{1,t-12} + \theta_{1,2}\,\tilde{\varepsilon}_{1,t-2} + \\ & + \theta_{1,1}\,\tilde{\varepsilon}_{1,t-1} + \tilde{\varepsilon}_{1,t}\,. \end{aligned} \tag{17}$$

The expanded form for the logarithm of sows farrowed at time t is

$$\begin{aligned}\log(z_{2,t}) = {} & -\phi_{2,2}\phi_{2,24}\log(z_{2,t-39}) + (\phi_{2,2} - \phi_{2,1})\phi_{2,24}\log(z_{2,t-38}) + (\phi_{2,1} - 1)\phi_{2,24}\log(z_{2,t-37}) + \\ & + (\phi_{2,2}\phi_{2,24} - \phi_{2,12}\phi_{2,2})\log(z_{2,t-27}) + \left((\phi_{2,1} - \phi_{2,2})\phi_{2,24} + \phi_{2,12}\phi_{2,2} - \phi_{2,1}\phi_{2,12}\right)\log(z_{2,t-26}) + \\ & + \left((1 - \phi_{2,1})\phi_{2,24} + (\phi_{2,1} - 1)\phi_{2,12}\right)\log(z_{2,t-25}) + (\phi_{2,12} - \phi_{2,24})\log(z_{2,t-24}) + \\ & + (\phi_{2,12} - 1)\phi_{2,2}\log(z_{2,t-15}) + \left((1 - \phi_{2,12})\phi_{2,2} + \phi_{2,1}\phi_{2,12} - \phi_{2,1}\right)\log(z_{2,t-14}) + \\ & + \phi_{2,24}\log(z_{2,t-36}) + \left((1 - \phi_{2,1})\phi_{2,12} + \phi_{2,1} - 1\right)\log(z_{2,t-13}) + (1 - \phi_{2,12})\log(z_{2,t-12}) + \\ & + \phi_{2,2}\log(z_{2,t-3}) + (\phi_{2,1} - \phi_{2,2})\log(z_{2,t-2}) + (1 - \phi_{2,1})\log(z_{2,t-1}) + \theta_{2,2}\theta_{2,24}\,\tilde{\varepsilon}_{2,t-26} + \\ & + \theta_{2,1}\theta_{2,24}\,\tilde{\varepsilon}_{2,t-25} + \theta_{2,24}\,\tilde{\varepsilon}_{2,t-24} + \theta_{2,12}\theta_{2,2}\,\tilde{\varepsilon}_{2,t-14} + \theta_{2,1}\theta_{2,12}\,\tilde{\varepsilon}_{2,t-13} + \theta_{2,12}\,\tilde{\varepsilon}_{2,t-12} + \theta_{2,2}\,\tilde{\varepsilon}_{2,t-2} + \\ & + \theta_{2,1}\,\tilde{\varepsilon}_{2,t-1} + \tilde{\varepsilon}_{2,t}\,.\end{aligned} \qquad (18)$$

The expanded form for the logarithm of the litter rate at time t can be summarized by using model-based estimates for monthly pig crop and monthly sows farrowed as

$$\log(\rho_t) = \log\left(\hat{z}_{1,t}^{(1)}\right) - \log\left(\hat{z}_{2,t}^{(1)}\right) + \tilde{\varepsilon}_{\rho,t}\,, \qquad (19)$$

so that the first equation in (15) is satisfied while optimizing the penalized likelihood for the model (17) and (18).

3.2 Model for quarterly estimates

Table 6: Notation used in Section 3.2.

Notation	Description
$z_{1,t}$	Monthly pig-crop at time t.
$y_{3,t}$	Quarterly inventory for the first weight group at time t.
$y_{4,t}$	Quarterly inventory for the second weight group at time t.
$y_{5,t}$	Quarterly inventory for the third weight group at time t.
$y_{6,t}$	Quarterly inventory for the fourth weight group at time t.
α_1	Percentage of pig-crop that do not transition from the first weight group to the second.
α_2	Percentage of pig-crop that do not transition from the second weight group to the third.
α_3	Percentage of pig-crop that do not transition from the third weight group to the fourth.
α_4	Percentage of pig-crop that do not transition from the fourth weight group to the slaughter facilities.
ζ_t	Survival rate associated with $y_{1,t}$, i.e. the monthly cohort of pig-crop born at time t.
$\varepsilon_{3,t}$	Statistical error in modeling the size of the first weight group of market hogs.
$\varepsilon_{4,t}$	Statistical error in modeling the size of the second weight group of market hogs.
$\varepsilon_{5,t}$	Statistical error in modeling the size of the third weight group of market hogs.
$\varepsilon_{6,t}$	Statistical error in modeling the size of the fourth weight group of market hogs.

Similar to the proposal of Pollard (1966), the equations governing the behavior of the weight classes are defined as:

$$\begin{cases} y_{3,t} = \zeta_{t-1}\, z_{1,t-1} + \zeta_{t-2}\, z_{1,t-2} + \zeta_{t-3}\, \alpha_1\, z_{1,t-3} + \varepsilon_{3,t}, \\ y_{4,t} = \zeta_{t-3}\, (1-\alpha_1)\, z_{1,t-3} + \zeta_{t-4}\, z_{1,t-4} + \zeta_{t-5}\, \alpha_2\, z_{1,t-5} + \varepsilon_{4,t}, \\ y_{5,t} = \zeta_{t-5}\, (1-\alpha_2)\, z_{1,t-5} + \zeta_{t-6}\, \alpha_3\, z_{1,t-6} + \varepsilon_{5,t}, \\ y_{6,t} = \zeta_{t-6}\, (1-\alpha_3)\, z_{1,t-6} + \zeta_{t-7}\, \alpha_4\, z_{1,t-7} + \varepsilon_{6,t}, \end{cases} \tag{20}$$

where $\alpha_i \in [0,1]$, for any $i = 1, \ldots, 4$, and the survival rate $\zeta_t \in (0,1]$ is associated with the monthly cohort $z_{1,t}$, such that the adjusted values of pig crop are propagated by accounting for pig losses within each cohort.

The relationships in (20) constrain the number of hogs in each weight group to be consistent with the number of piglets born in the past that are still alive. This formulation provides an estimate of the survival probabilities of each monthly cohort during its lifespan. The simplified

system of equations (20) can be extended by considering additional effects from lagged residuals, including nonlinear terms. However, since this model has a high number of parameters, the contribution of additional terms is not considered here.

The model (20) can track changes in the monthly pig cohorts. Since the survival rates are cohort dependent, they are restricted during the estimation phase by optimizing the lagged differences. This approach has been inspired by the use of penalties as formulated in the P-spline proposal of Eilers and Marx (1996), which were shown to be computationally efficient by maintaining the model as elementary as possible without over-fitting the data. This technique provides smooth survival rates that quickly adapt by accounting for the temporal evolution of the hog population. Any type of death is considered and summarized into a single monthly value, which represents an overall estimate of the percentage of monthly pig-crop that reaches the proper weight to be slaughtered. For example, a cohort of piglets born during month t has a unique survival rate that quantifies its chances to reach market maturity. This survival rate ζ_t is well-localized in time. Survival rates may be low during epidemic periods. In such cases, a specific cohort born during the month t would experience a high mortality rate. High values of survival rates denote periods that are not affected by systemic shocks.

4 Estimation

Table 7: Notation used in Section 4 (same notation in Table 1).

Notation	Description
$y_{k,t}^{(1)}$	Estimate for variable k at time t produced during the pre-processing stage.
$y_{k,t}^{(2)}$	Estimate for variable k at time t produced by the pre-board.
$\hat{\mathbf{y}}_t^{(1)}$	Vector of model-based estimates at time t produced by the first estimation stage.
$\mathbf{y}_t^{(2)}$	Vector of update estimates at time t provided by the pre-board.
$\hat{\mathbf{y}}_t^{(2)}$	Vector of model-based estimates at time t produced by the second estimation stage.

Estimation of the model parameters occurs in two stages (see Figure 1). The aim of the first stage is to produce estimates that combine the dynamic history, the survey data, and the state recommendations. These estimates, $\hat{\mathbf{y}}_t^{(1)}$, will be then evaluated by the pre-board that produces a set of preliminary estimates, $\mathbf{y}_t^{(2)}$, for the variables of interest. The second estimation stage uses the preliminary estimates provided by the pre-board, $\mathbf{y}_t^{(2)}$, to estimate the time series models to produce estimates, $\hat{\mathbf{y}}_t^{(2)}$, for ASB evaluation.

The maximum likelihood methodology can be adapted to develop a procedure capable of handling time series with different time resolutions. The proposed time series methodology consists of two algorithms that respectively produce estimates for inventory items (i.e. for variable $k = 3, \dots, 7$) and non-inventory items (i.e. for variable $k = 1,2$). The same methodology is also applied during the second estimation round using the updated dataset, where the information provided by the survey and state recommendations, $y_{k,t}^{(1)}$, is replaced by the pre-board estimates $y_{k,t}^{(2)}$.

Both algorithms used for the estimation of the model parameters are iterative in nature and take advantage of other methods proposed for solving nonlinear optimization problems. Generally speaking, an iterative algorithm is used when there is no closed-form solution to an optimization problem. Thus, one starts from an initial guess for each of the parameter values, which is updated using the method of steepest descent. These adjustments produce a better set of estimates resulting in a smaller sum of the squared residuals. This process is repeated until no further improvements are possible. In particular, the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (Fletcher 1987) is used by the iterative procedure that optimizes the parameter estimates and updates the residuals of the equations in (16). In addition, the limited memory algorithm for bound constrained optimization (Byrd et al. 1995) is applied for optimization of the parameters in the system of equations (20).

4.1 Optimization for monthly estimates

Table 8: Notation used in Section 4.1.

Notation	Description
s	Index denoting the estimation stage. E.g. $s = 1$ denotes the first estimation stage as shown in Figure 1.
$y_{k,t}^{(s)}$	Estimate for variable k at time t produced during the pre-processing stage, if $s = 1$, or by the pre-board, if $s = 2$.
t_0	Current estimation time.
$\tilde{\varepsilon}_{1,t}$	Statistical error in modeling the logarithm of pig crop $\log(y_{1,t})$.
$\tilde{\varepsilon}_{2,t}$	Statistical error in modeling the logarithm of sows farrowed $\log(y_{2,t})$.
$\tilde{\varepsilon}_{\rho,t}$	Statistical error in modeling the logarithm of litter rate $\log(\rho_t)$.
δ	A non-negative scalar denoting the LASSO penalty.
γ	A non-negative penalty that governs the importance of the litter rate.
φ	Parameter controlling for the farrowing rate.
$\phi_{k,12^i j}$	Parameters controlling for the auto-regressive model of variable k. For $i = 0,1$, and $j = 1,2$, this notation denotes the parameters $\phi_{k,1}$, $\phi_{k,2}$, $\phi_{k,12}$, and $\phi_{k,24}$.

$\theta_{k,12^i j}$ Parameters controlling for the moving average model of variable k. For $i = 0,1$, and $j = 1,2$, this notation denotes the parameters $\theta_{k,1}$, $\theta_{k,2}$, $\theta_{k,12}$, and $\theta_{k,24}$.

To optimize the total loss associated with estimation of the parameters in (16), the following quantity is minimized:

$$\sum_{t=1}^{t_0}\left[\sum_{k=1}^{2}\left(\tilde{\varepsilon}_{k,t}\right)^2+\gamma\left(\tilde{\varepsilon}_{\rho,t}\right)^2\right]+\sum_{\ell=0}\left(z_{2,t-3\ell+2}^{(s)}-\varphi\, y_{7,t-3\ell}^{(s)}\right)^2$$
$$+\delta\sum_{k=1}^{2}\sum_{i=0}^{1}\sum_{j=1}^{2}\left(|\phi_{k,12^i j}|+|\theta_{k,12^i j}|\right), \qquad (21)$$

where the parameters $\phi_{k,12^i j}$ and $\theta_{k,12^i j}$, for $k = 1,2$; $i = 0,1$; and $j = 1,2$, have an impact on the residuals $\tilde{\varepsilon}_{k,t}$ and $\tilde{\varepsilon}_{\rho,t}$. The non-negative scalar γ governs the importance of the litter rate, and it can be found by applying standard cross-validation methods (Roberts and Nowak 2014). The penalty δ is used to perform LASSO regression (Tibshirani 1996) on the time series models in (17), (18) and (19). By minimizing (21), LASSO regression is simultaneously performed on the three equations in (16), and it also accounts for the expressions of the system (15). This makes it possible to avoid separate estimation of the equations formulated in the system (16), because a unified procedure accounts for inter-relationships that affect the behavior of other variables.

Solution for the problem stated in (21) requires a numerical procedure to compute suitable parameter estimates. The initial choice of values for this iterative estimation algorithm is $\varphi = \frac{1}{6}$, which is approximately the proportion of sows farrowed in a month from the breeding herd. The time series parameters are set to $\phi_{k,12^i j} = 0$, and $\theta_{k,12^i j} = 0$, for $k = 1,2$; $i = 0,1$; and $j = 1,2$, which reflect the equilibrium of a static process. The initial values of the residuals $\tilde{\varepsilon}_{k,t}$ and $\tilde{\varepsilon}_{\rho,t}$ are also set to zero, and updated at each iteration. Problems of failure to converge or convergence to a local minimum are avoided by minimizing the quantity in (21), since this process is equivalent to the minimization of a convex function, which has a unique solution.

The optimization of the quantity in (21) is conducted for each value of δ in the set $\{0.8^i : i = 0, \ldots, 40\}$ by performing the following steps:

- For a given set of values for the parameters and the residuals, perform one updating step of the BFGS algorithm to produce better values for the parameters, such that the sum of squared residuals in (21) becomes smaller;

- Given the new values of the parameter, produce new values of residuals;
- Determine whether the convergence is achieved. If not, repeat step 1 and 2 until convergence.

Once parameter estimates are produced for the specified values of the penalty δ, model selection is performed by setting to zero those values that, overall, are not significantly different from zero. The same regression mechanism (as explained in the previous three iterative steps) is executed for fitting the model by setting $\delta = 0$. Thus, the parameters are freely allowed to vary without imposing any penalty during the optimization, but those forced to zero automatically exclude variables that are not closely associated with the parameters to be estimated.

Parameters are determined to be non-significant by a voting system (of 41 votes) that counts how many times a parameter is significantly different from zero. The voting system is based on the trajectory formed by the estimates of a parameter that are obtained for different values of δ. In particular, the sequence of parameter estimates is processed by evaluating the difference of consecutive penalties computed as

$$\sum_{k=1}^{2}\sum_{i=0}^{1}\sum_{j=1}^{2}\left(|\phi_{k,12^{i}j}| + |\theta_{k,12^{i}j}|\right). \qquad (22)$$

The fitted values for variables $k = 1,2$ at time t_0, $t_0 - 1$ and $t_0 - 2$ form the estimates on a quarterly basis. The ratio between the estimate for pig crop and sows farrowed on the current quarter form the estimates of pigs per litter (litter rate). The estimate for the breeding herd, $\widehat{y}_{7,t_0}^{(s)}$, is obtained by using the forecast of the monthly sows farrowed two months ahead in the objective function, so that $\widehat{y}_{7,t_0}^{(s)} = \hat{z}_{2,t_0+2}^{(s)} / \widehat{\varphi}$. This formulation is derived from (15) and accounts for the biological gestation time (about three months). Recall that monthly sows farrowed z_{2,t_0}, z_{2,t_0+1}, and z_{2,t_0+2} sum up to form the value y_{2,t_0+3} for the next quarter (see NASS's Quarterly Hogs and Pigs Report, reports from December 2018 and March 2019 are in the appendix of Chapter 2).

By estimating the parameters controlling the value of the quantity (21), it is possible to obtain fitted values for monthly pig crop $\hat{z}_{1,t_0-1}^{(s)}$, $\hat{z}_{1,t_0-2}^{(s)}$, $\hat{z}_{1,t_0-3}^{(s)}$, and sow farrowed $\hat{z}_{2,t_0-1}^{(s)}$, $\hat{z}_{2,t_0-2}^{(s)}$, $\hat{z}_{2,t_0-3}^{(s)}$. These fitted values form the quarterly estimates for pig crop, $\widehat{y}_{1,t_0}^{(s)}$, and sows farrowed, $\widehat{y}_{2,t_0}^{(s)}$, as formulated in equation (4).

This algorithm is also used to process the data to be used in the second estimation stage. As explained earlier, the second estimation stage incorporates the new values of historical and adjusted statistics after the pre-board sets their updated values for the current and past four quarters.

4.2 Optimization for quarterly estimates

Table 9: Notation used in Section 4.2.

Notation	Description
$\widehat{y}_{3,t}^{(s)}$	Quarterly inventory estimate for the first weight group at time t (output of the s estimation stage).
$\widehat{y}_{4,t}^{(s)}$	Quarterly inventory estimate for the second weight group at time t (output of the s estimation stage).
$\widehat{y}_{5,t}^{(s)}$	Quarterly inventory estimate for the third weight group at time t (output of the s estimation stage).
$\widehat{y}_{6,t}^{(s)}$	Quarterly inventory estimate for the fourth weight group at time t (output of the s estimation stage).
α_1	Percentage of pig-crop that do not transition from the first weight group to the second.
α_2	Percentage of pig-crop that do not transition from the second weight group to the third.
α_3	Percentage of pig-crop that do not transition from the third weight group to the fourth.
α_4	Percentage of pig-crop that do not transition from the fourth weight group to the slaughter facilities.
ζ_t	Survival rate associated with $y_{1,t}$, i.e. the monthly cohort of pig-crop born at time t.
∇^d	Difference operator of order d at lag $S = 1$. E.g. the notation $\nabla^3\zeta_t$ is equivalent to $\zeta_t - 3\,\zeta_{t-1} + 3\,\zeta_{t-2} - \zeta_{t-3}$.
$\varepsilon_{3,t}$	Statistical error in modeling the size of the first weight group of market hogs.
$\varepsilon_{4,t}$	Statistical error in modeling the size of the second weight group of market hogs.
$\varepsilon_{5,t}$	Statistical error in modeling the size of the third weight group of market hogs.
$\varepsilon_{6,t}$	Statistical error in modeling the size of the fourth weight group of market hogs.
t_0	Current estimation time, which is equivalent to the time length of the data period used for the estimation of the model parameters.

To reduce the sum of squared residuals associated with estimation of the parameters in (20), the following quantity is minimized:

$$\frac{1}{t_0}\sum_{t=1}^{t_0}\sum_{k=3}^{6}\left(\varepsilon_{k,t}\right)^2+\sum_{t=1}^{t_0}\left(|\zeta_t-1|+\sum_{d=1}^{3}|\nabla^d\zeta_t|\right), \qquad (23)$$

such that $\alpha_i, \zeta_t \in [0,1]$, for any $i = 1, \dots, 4$, and $t \in \mathbb{Z}$. All parameters α_i and ζ_t govern the behavior of the residuals $\varepsilon_{k,t}$.

For the equations in (20), the initial choice of the parameters α_i is set to 0.25, for $i = 1,2$, and 0.75 for $i = 3,4$. These values are based on the growth rates studied by Shull (2013) (Table 20, page 114), such that the life-span of a single market hog is consistent with the expected growth of its monthly cohort with respect to the four weight groups. At the same time, the initial values of the survival rates ζ_t are set to 1, for any $t \in \mathbb{Z}$, so as to represent a hypothetical world without disease outbreaks. These values, however, will be dynamically updated to reflect the effective status of the hog population.

The algorithm proposed by Byrd et al. (1995) allows for simultaneous minimization of the quantity (23) with respect to the parameters involved in the system of equations (20). This approach guarantees that the final results satisfy the bound constraints set by the model. Under the assumption that the dataset used for regression provides enough evidence that reflects the true status of the swine population, the proposed methodology, due to its flexibility, is able to quickly adapt and provide improved estimates in the event of a systemic shock (see section 5 for the performance evaluation of this model).

This same algorithm is applied for the second estimation stage. In general, the estimates produced for stage $s = 1,2$, are provided by the estimated model parameters for $\widehat{y}_{k,t}^{(s)}$, where $k = 3, \dots, 6$.

4.3 Updating the dataset

Table 10: Notation used in Section 4.3.

Notation	Description
$\mathbf{y}_{t_0}^{(1)}$	Adjusted survey estimates at time t_0 use to fit the model.
$\mathbf{y}_{t_0}^{(2)}$	Estimates set by the pre-board at time t_0.
t_0	Current estimation time.
$\hat{z}_{1,t}^{(1)}$	Model-based estimate of monthly pig crop at time t.
$\hat{z}_{2,t}^{(1)}$	Model-based estimate of monthly sows farrowed at time t.
$y_{1,t_0}^{(2)}$	Pre-board value for quarterly pig crop at time t_0.
$y_{2,t_0}^{(2)}$	Pre-board value for quarterly sows farrowed at time t_0.
$z_{1,t}^{(2)}$	Calibrated monthly pig crop at time t.
$z_{2,t}^{(2)}$	Calibrated monthly sows farrowed at time t.

The second estimation stage is performed after the information update step, which is between the two estimation stages in Figure 1. As discussed next, the historical information requires minimal adjustments to be more consistent with the most recent administrative data (such as the weekly slaughter data that were not available when the historical values were initially set). The adjusted survey estimates, $\mathbf{y}_{t_0}^{(1)}$, that are used in the time series analysis are also modified to reflect the state of the dynamic systems in (15) and (20) by incorporating other pieces of information available to the commodity experts.

The members of the pre-board provide a set of estimates, $\mathbf{y}_{t_0}^{(2)}$, for the current time t_0. These estimates are produced only for quarterly summary statistics; therefore, the monthly estimates of pig crop and sows farrowed from the first estimation stage, $\hat{z}_{k,t_0-h}^{(1)}$, for $h \in \{1,2,3\}$ and $k = 1,2$, are calibrated to match pre-board quarterly estimates. Similar to the process described in section 2.3, the calibrated values for monthly pig crop, $z_{1,t}^{(2)}$, and sows farrowed, $z_{2,t}^{(2)}$ can be expressed as

$$\begin{cases} z_{1,t}^{(2)} = \hat{z}_{1,t}^{(1)} + \left(\hat{z}_{1,t}^{(1)}\right)^2 \dfrac{y_{1,t_0}^{(2)} - \sum_{i=t_0-2}^{t_0} \hat{z}_{1,i}^{(1)}}{\sum_{i=t_0-2}^{t_0} \left(\hat{z}_{1,i}^{(1)}\right)^2}, \\ z_{2,t}^{(2)} = \hat{z}_{2,t}^{(1)} + \left(\hat{z}_{2,t}^{(1)}\right)^2 \dfrac{y_{2,t_0}^{(2)} - \sum_{i=t_0-2}^{t_0} \hat{z}_{2,i}^{(1)}}{\sum_{i=t_0-2}^{t_0} \left(\hat{z}_{2,i}^{(1)}\right)^2}, \end{cases} \tag{24}$$

where $t \in \{t_0, t_0 - 1, t_0 - 2\}$.

The calibrated values for pig crop and sows farrowed at the monthly level together with the quarterly values produced by the pre-board replace the survey estimates used in the dataset for the first estimation stage. The second estimation stage is performed on the updated information that accounts for additional expert knowledge. The time series algorithms are performed as explained in Section 4.1 and 4.2.

5 Data Analyses

Table 11: Notation used in Section 5.

Notation	Description
$\widehat{y}_{k,t}^{(2)}$	Estimate for variable k at time t produced by the second estimation stage.
$y_{k,t}^{*}$	True value for variable k at time t
MAE_k	Mean absolute error produced for variable k.
RMSE_k	Root mean square error produced for variable k.
MAPE_k	Mean absolute percentage error produced for variable k.
MPE_k	Mean percentage error produced for variable k.
T	Time length of the data period used for the evaluation of the proposed model.

The constrained state-space model developed by Busselberg (2013) (called the KFM as in Chapter 4) and the sequential generalized linear models suggested by Kedem and Pan (2015) are currently used at NASS to produce quarterly estimates for total hogs, breeding herd, the inventory numbers for the four weight classes, pig crop, sows farrowed and litter rate. However, only the results for the KFM are compared to the results of the new model. The results from the model proposed by Kedem and Pan (2015) do not satisfy the basic constraints.

The proposed model is compared with the KFM discussed in Chapter 4 based on classical model selection criteria from the machine learning community, where models are usually over-parameterized. Hyndman and Koehler (2006) provide a detailed review about measures of accuracy.

In general, the criteria adopted to compare the performance of a regression model include (but are not restricted to) the following measures (Hyndman and Koehler 2006; Khan and Hildreth 2003):

- **Mean absolute error** (MAE) is calculated by taking the arithmetic average of absolute residuals, which are computed as the difference between the predicted value $\widehat{y}_{k,t}^{(2)}$, and true value $y_{k,t}^{*}$:

$$\mathrm{MAE}_k = \frac{1}{T}\sum_{t=1}^{T}\left|y_{k,t}^{*} - \widehat{y}_{k,t}^{(2)}\right|. \qquad (25)$$

  MAE reports the magnitude of the residuals, and it is robust to outliers.
- **Root mean square error** (RMSE) is very similar to MAE, but it is computed as the square root taken over the average of the squared residuals:

$$\mathrm{RMSE}_k = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(y_{k,t}^{*} - \widehat{y}_{k,t}^{(2)}\right)^2}. \qquad (26)$$

  In comparison to the MAE, RMSE uses quadratic residuals to emphasize the presence of outliers.
- **Mean absolute percentage error** (MAPE) is defined by scaling the absolute residuals with respect to the true value:

$$\mathrm{MAPE}_k = \frac{100\%}{T}\sum_{t=1}^{T}\left|\frac{y_{k,t}^{*} - \widehat{y}_{k,t}^{(2)}}{y_{k,t}^{*}}\right|. \qquad (27)$$

  This index reports the relative distance between predictions and true values as a percentage. MAPE is also robust to outliers as MAE.
- **Mean percentage error** (MPE) is computed as:

$$\mathrm{MPE}_k = \frac{100\%}{T}\sum_{t=1}^{T}\left(\frac{y_{k,t}^{*} - \widehat{y}_{k,t}^{(2)}}{y_{k,t}^{*}}\right). \qquad (28)$$

  This measure indicates whether the model is underestimating the true values (by having more negative residuals), or is overestimating (by having more positive residuals).

First, the data from 2013 to 2017 are used for comparing the estimates produced by the two models versus the initial board estimates, $\hat{\mathbf{y}}_t$, and the final estimates obtained after several board revisions. NASS historical estimates have been used for this analysis starting from the first quarter in 2008. Quarterly estimates for pig crop, sows farrowed, breeding herd, and the four weight groups are produced directly from the models. Total market hogs are computed by aggregating the inventory estimates $\hat{y}_{k,t}^{(2)}$, for $k = 3, \ldots, 6$. Total hogs in the US are computed by adding the number of breeding sows and boars to the total value of market hogs.

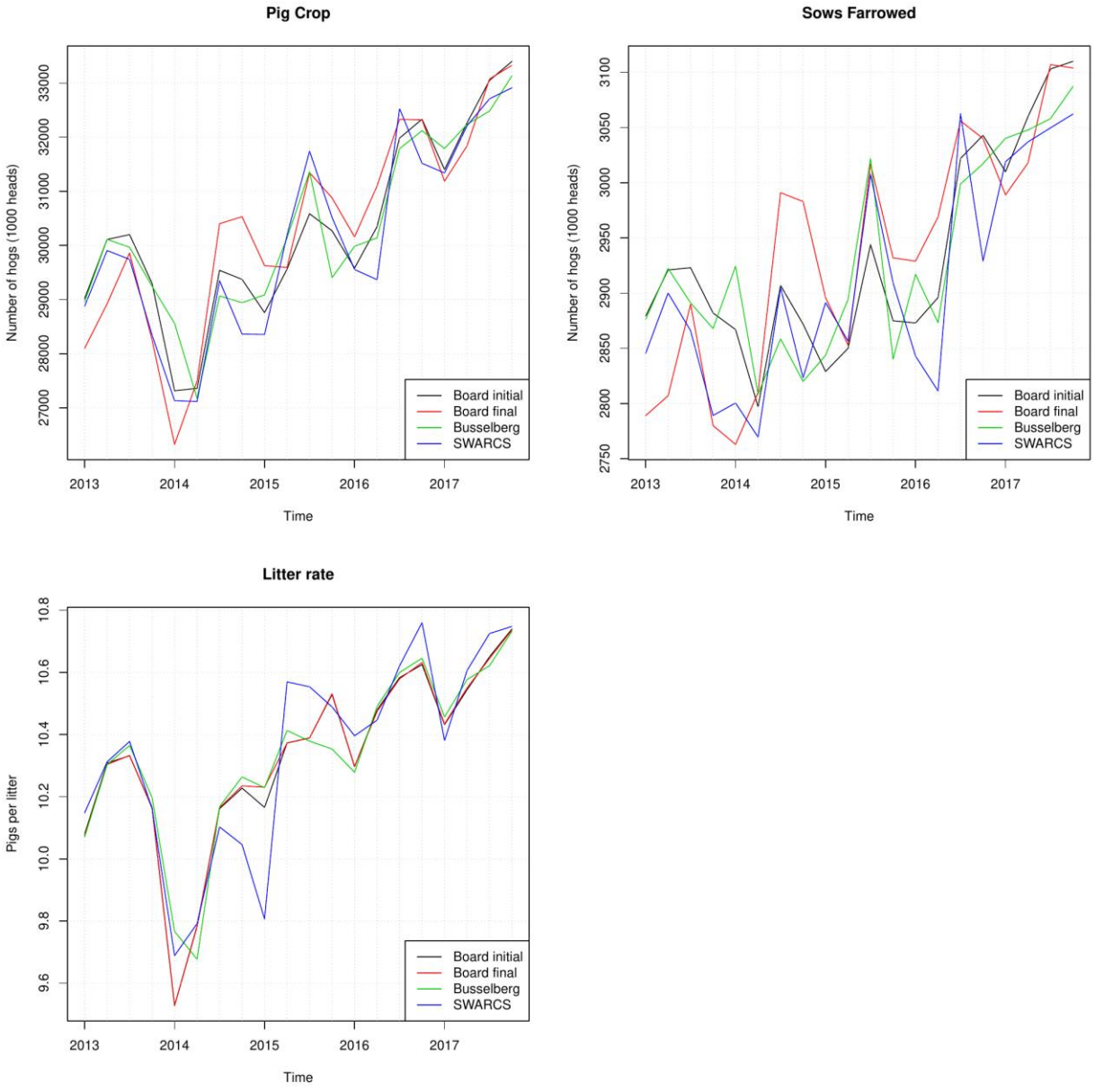

Figure 4: Quarterly estimates based on monthly data from 2013 to 2017 in US.

In Figure 4, initial official and final estimates from the ASB are compared to those from the two models from March 2013 to December 2017. Each graph in Figure 4 has a generally increasing trend with a notable shock between 2013 and 2015. In this case, the shock was caused by Porcine Epidemic Diarrhea virus (PEDv). PEDv is a highly contagious coronavirus that attacks hogs of all ages, but is particularly deadly to suckling pigs that have not weaned. In each graph, the board final (red line) reaches its lowest value at 2014. All other estimates are low around 2014 as well, but looking at the board final compared to the other estimates highlights the extent of the shock. In other words, the initial, KFM, and SWARCS all reacted to the shock and produced low estimates in pig crop, sows farrowed, and litter rate at or around 2014, but they were unable to account for the extent of lost. After revisions and more information was collected, the ASB was able to better account for those losses and revised the official estimates to reflect what is seen as the board final in Figure 4. After 2015, pig crop estimates stabilize fairly quickly, but sows farrowed and litter rate take a little more time to stabilize. A number of things could contribute to the erratic sows farrowed estimates after 2015, but it is most likely due to the industries response to PEDv as they try to increase or decrease the number of sows farrowed in response to the impact (or lack thereof) of PEDv on their stock. Litter rate for the most part stabilizes quickly after the shock but the estimates from SWARCS take about another year to tighten up with other estimates.

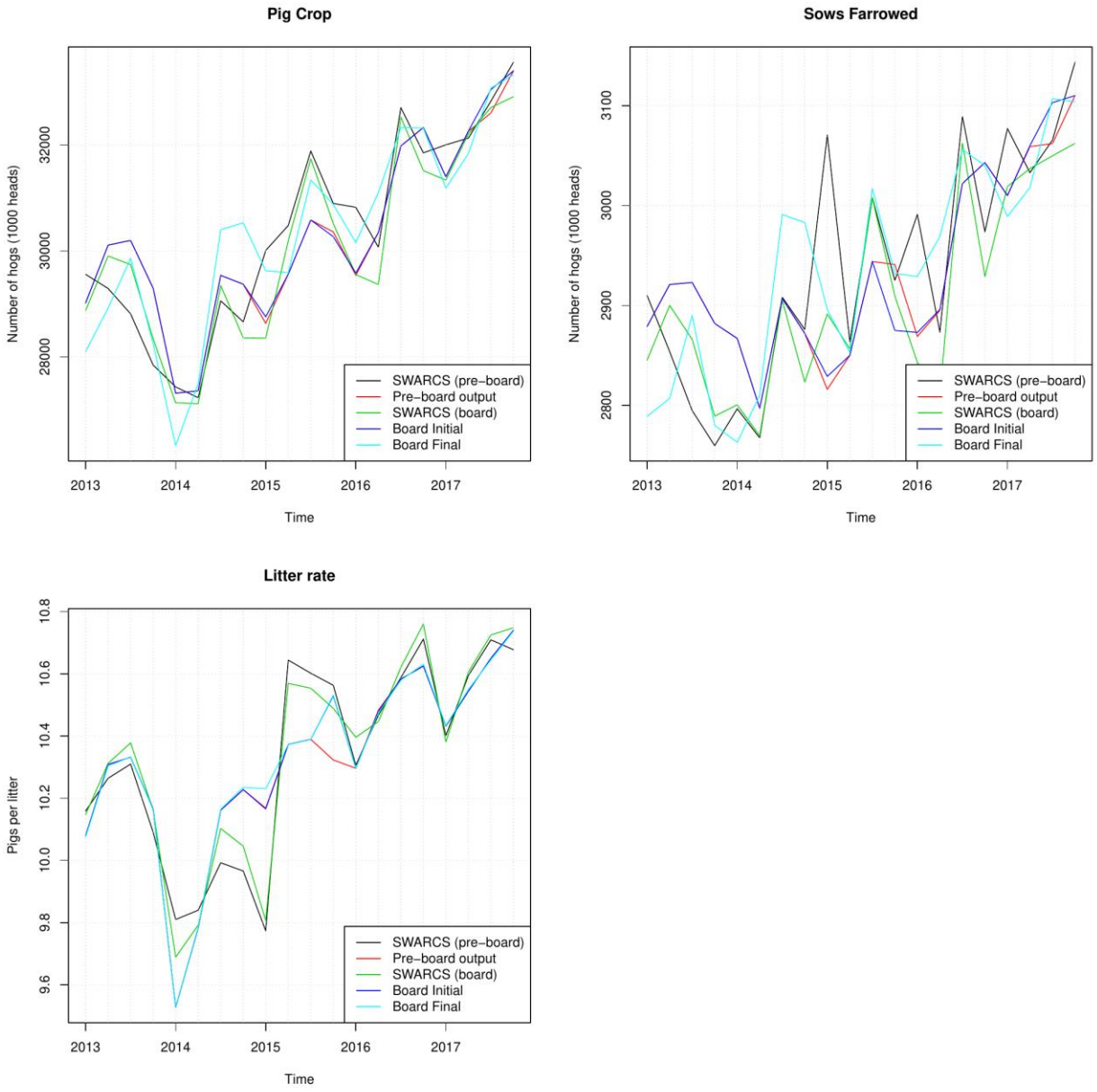

Figure 5: Sequential updates of quarterly estimates based on monthly data from 2013 to 2017 in US.

The evolution of the estimates from the earlier SWARCS estimates produced for the pre-board through the estimates after the final 5-years revision based on the 2017 US Census of Agriculture is displayed in Figure 5. The ASB's initial official estimates correspond closely with the pre-board estimates for most of the reported quarters. However, the final estimates are often quite different from the initial official estimates. Sometimes, but not always, the SWARCS model provides estimates for the pre-board that are closer to the final board estimates than the initial official estimates. Although monthly error components are incorporated in the

SWARCS model to improve the quarterly estimates of pig crop sows farrowed, the results are not encouraging.

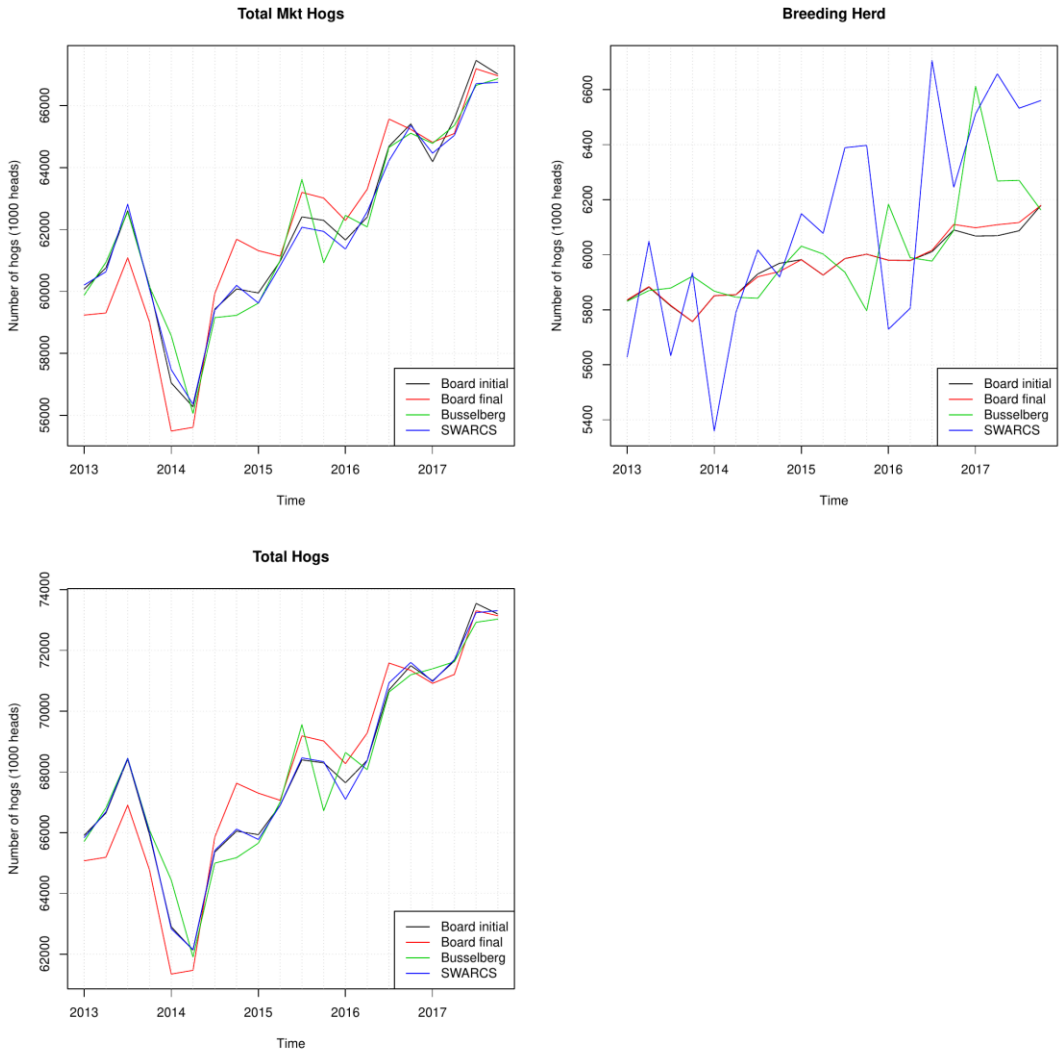

Figure 6: Quarterly estimates based on quarterly data from 2013 to 2017 in US.

Figure 6 shows the national total inventory, and its decomposition into total market hogs and breeding herd for each quarter from 2013 to 2017. The SWARCS model better estimates the number of total hogs, and it is superior to those from the KFM model during anomalous periods caused by disease outbreaks (e.g. see the estimates between December 2013 and June 2014). The SWARCS model is also closer to the initial board estimates than to the final estimates, and it is closer to the final estimates than the KFM. A very different behavior is observed when estimating the size of the breeding herd. The KFM is capable of producing better estimates for breeding herd, even though some estimates are too far from the final estimates to be useful. On the other hand, the SWARCS model is capable of capturing the underlying trend of the time series produced by the board, but its estimates of the breeding herd are highly variable, which

makes them less reliable. Perhaps this phenomenon can be mitigated by introducing a further dynamic equation to stabilize the model-based estimates of the breeding herd.

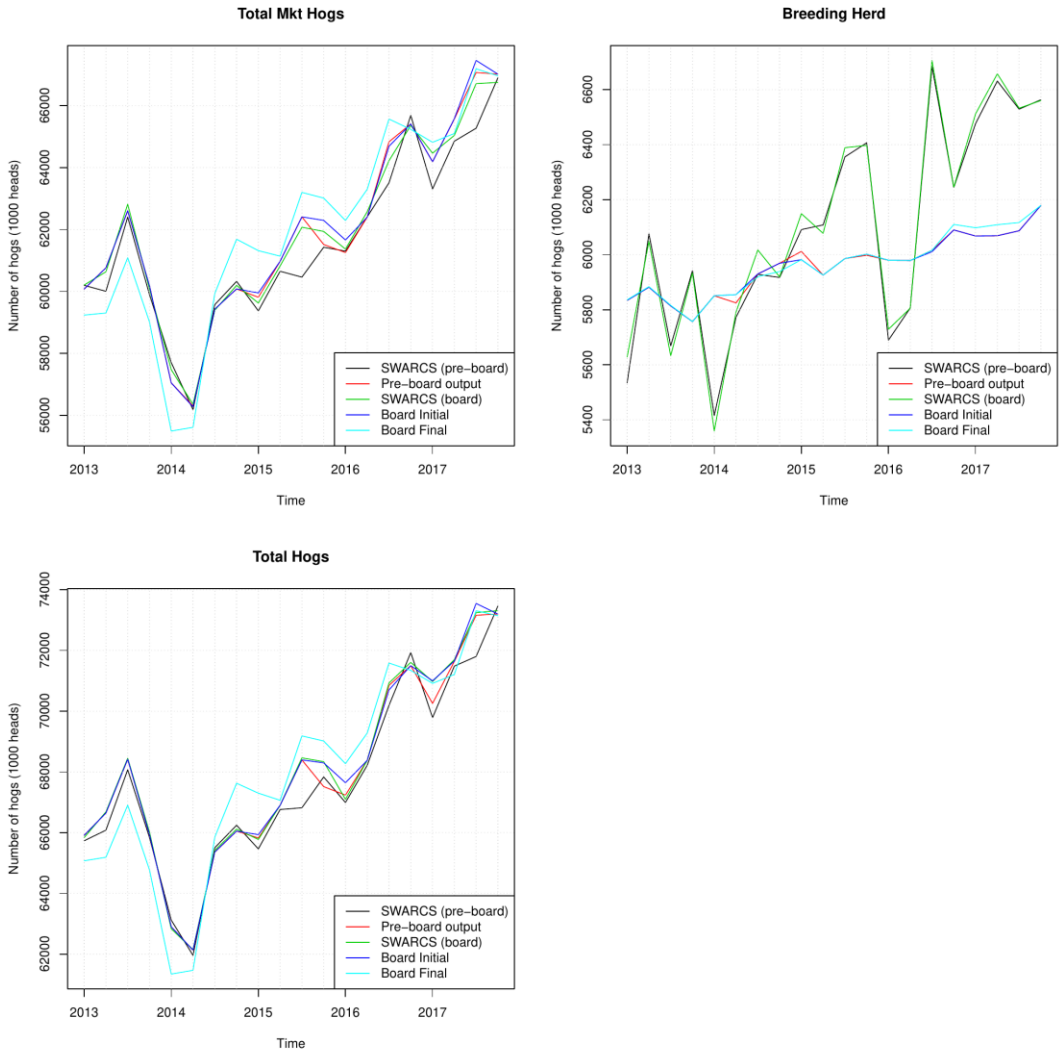

Figure 7: A comparison of the quarterly estimates for total market hogs, breeding herd, and total hog inventory from 2013 to 2017 in US.

The estimates of total market hogs, breeding herd, and total hog inventory are compared in Figure 7. The SWARCS estimates for breeding herd are quite volatile compared to the pre-board and final estimates, which in contrast are relatively close to each other. In contrast, the SWARCS estimates for total inventories are close to the pre-board and board estimates, which is likely due to the calibration process that moves the results towards the state recommended estimates. Even when the SWARCS estimates of total market hogs and total hogs inventories are not close to the final estimates, the model is mostly able to capture the underlying dynamics (e.g. local temporal trend and seasonality).

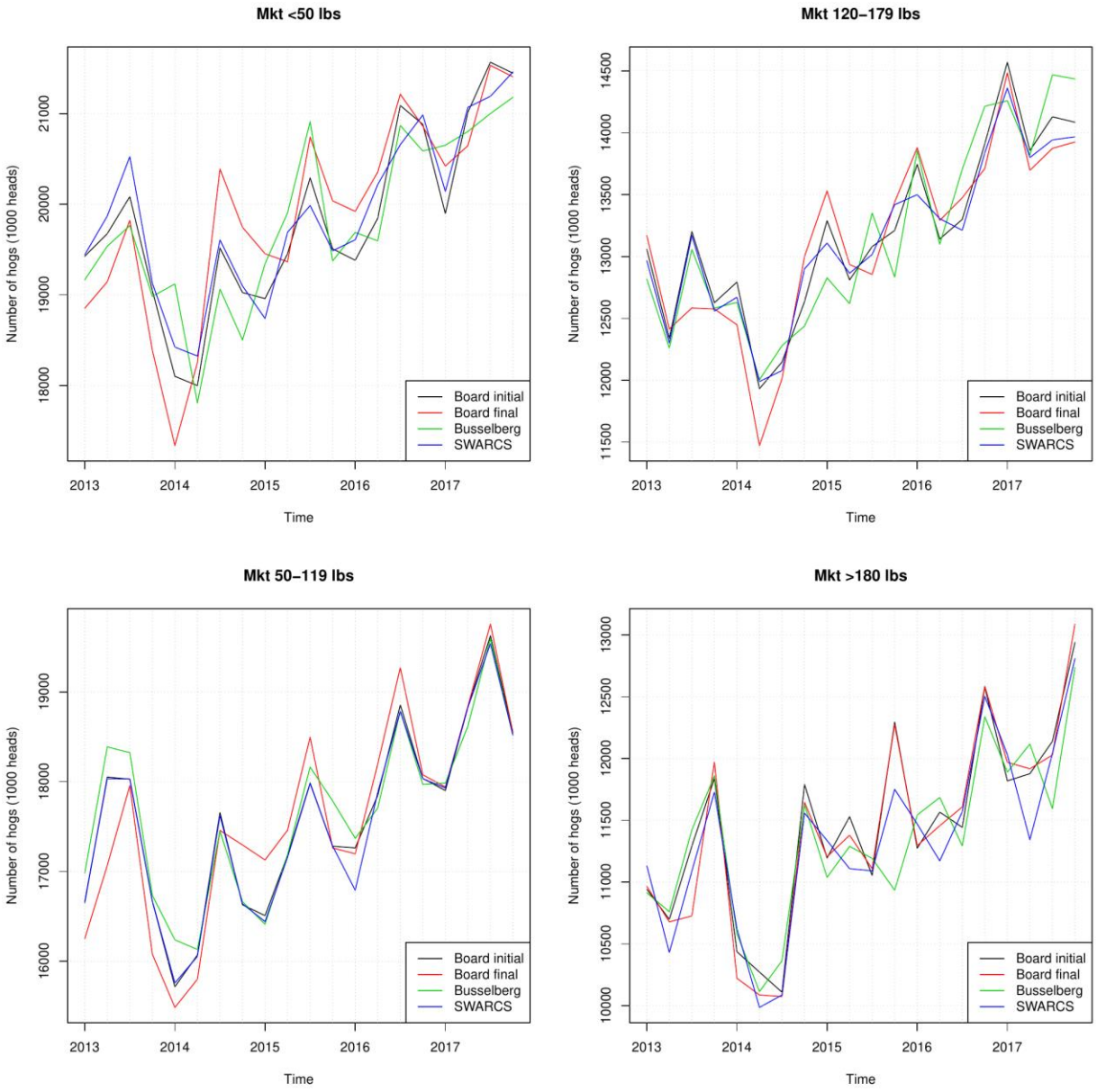

Figure 8: Market hogs distinct by weight classes from 2013 to 2017 in US.

The national inventories for the four weight groups are shown in Figure 8. The estimates produced by the two models are compared for each quarter from 2013 to 2017. The SWARCS model is able to approximately reproduce the initial board estimates for all the weight groups considered. Both models estimate the number of hogs between 50 lbs and 119 lbs accurately. The KFM becomes less accurate for the group of hogs weighing less than 50 lbs and for the two highest weight groups (120 to 179 lbs. and at least 180 lbs.)

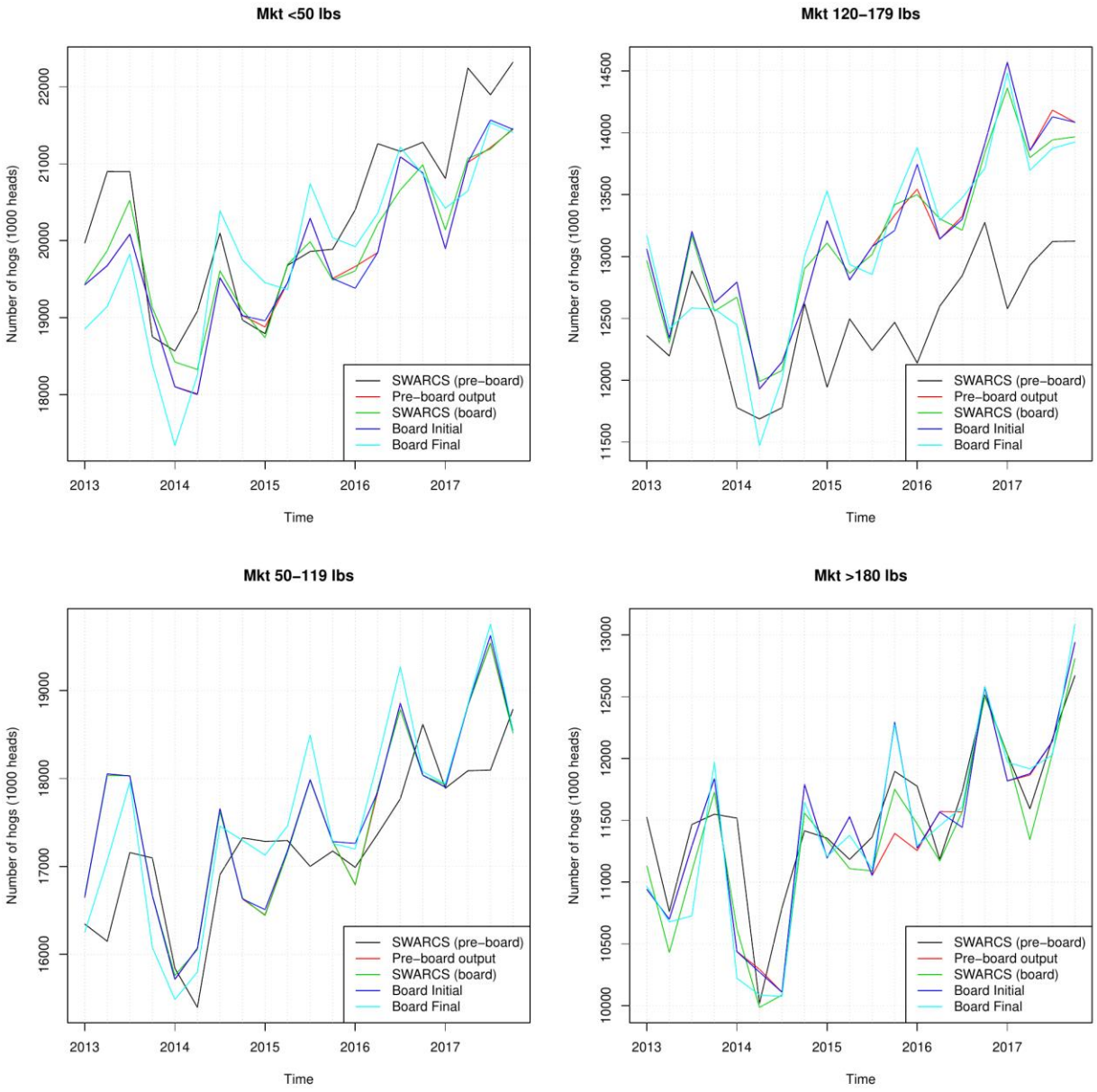

Figure 9: Sequential updates of market hogs distinct by weight classes from 2013 to 2017 in US.

Figure 9 shows the estimates produced for the inventories of the four weight groups. The SWARCS estimates produced for the pre-board are not as close as desired to either the final or the initial board estimates. However, the final model is able to provide board estimates that are close to the pre-board and initial board estimates. The SWARCS estimates produced for the pre-board can be improved by considering the state recommended estimates for these four groups. These values are not available for all 20 quarters used in this analysis; therefore, calibration has been conducted to capture the information in the state-recommended estimates for the four

weight groups. Better estimates should result if all currently available information for modeling is included in the SWARCS model.

Further comparisons between the SWARCS and the KFM are shown in Figures 10 and 11. These graphics shows the statistics as formulated in the equations (25), (26), (27) and (28). On average when compared to the initial official estimates, the KFM produced better estimates for pig crop, sows farrowed and breeding herd, while the SWARCS model was capable of producing better estimates for four inventory items and their totals. However, when comparing the results of the models with the final official estimates, with the exception of breeding herd, the SWARCS model tended to produce better estimates.

6 Future work and improvements

In this chapter, the biological growth of hogs from newborn piglets to market weight and the resulting numbers of hogs in various categories are accounted for by modeling both growth and survival rates under different conditions (e.g. presence/absence of disease outbreaks). Because only national estimates (and not state estimates) are produced, an initial signal of a disease outbreak within one or a few states may be masked.

The proposed model can be extended to produce state-level estimates that account for interstate transport. The quarterly survey does not provide this information, but other governmental sources may provide the in-flow and out-flow of hogs among the states. By adopting a dynamic graphical model at the state level, with the proper considerations made for the national level, more reliable model-based estimates can be produced.

A web scraping technique to detect disease outbreaks has been recently developed at NASS for making the model more flexible to systemic shocks (see the appendix). However, it is not clear how to include web-scraped information in the dataset adopted for the time series models. The current state of this technology provides warnings related to disease outbreaks affecting the hog population.

Further improvements can be made by accounting for the quality of survey data. Other improvements should consider the most recent development of imputation techniques for the quarterly swine survey. In addition, imputation techniques can benefit from the new model, by considering the dynamic of a herd within each single operation.

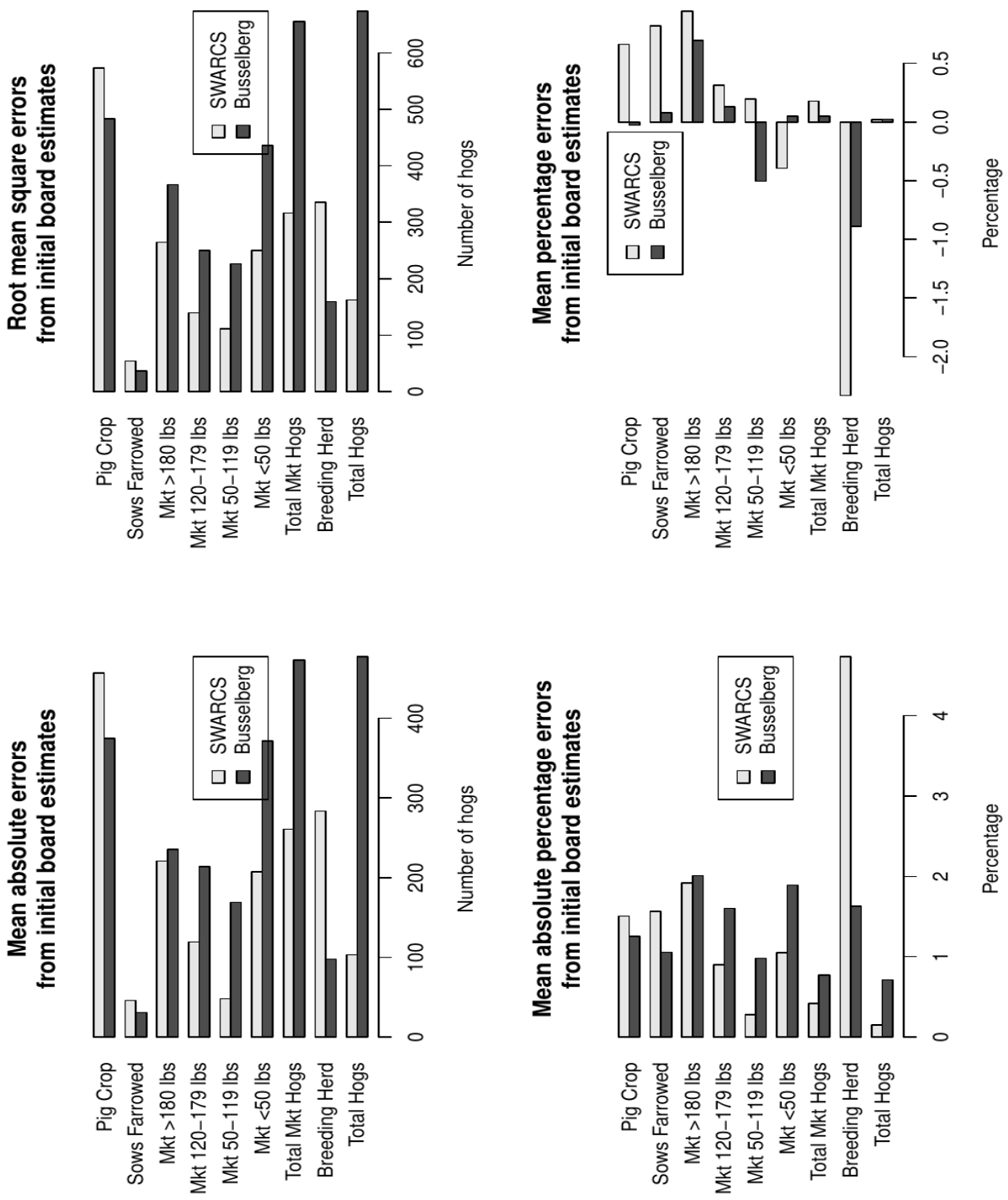

Figure 10: Prediction accuracy of the models when compared to the initial board estimates.

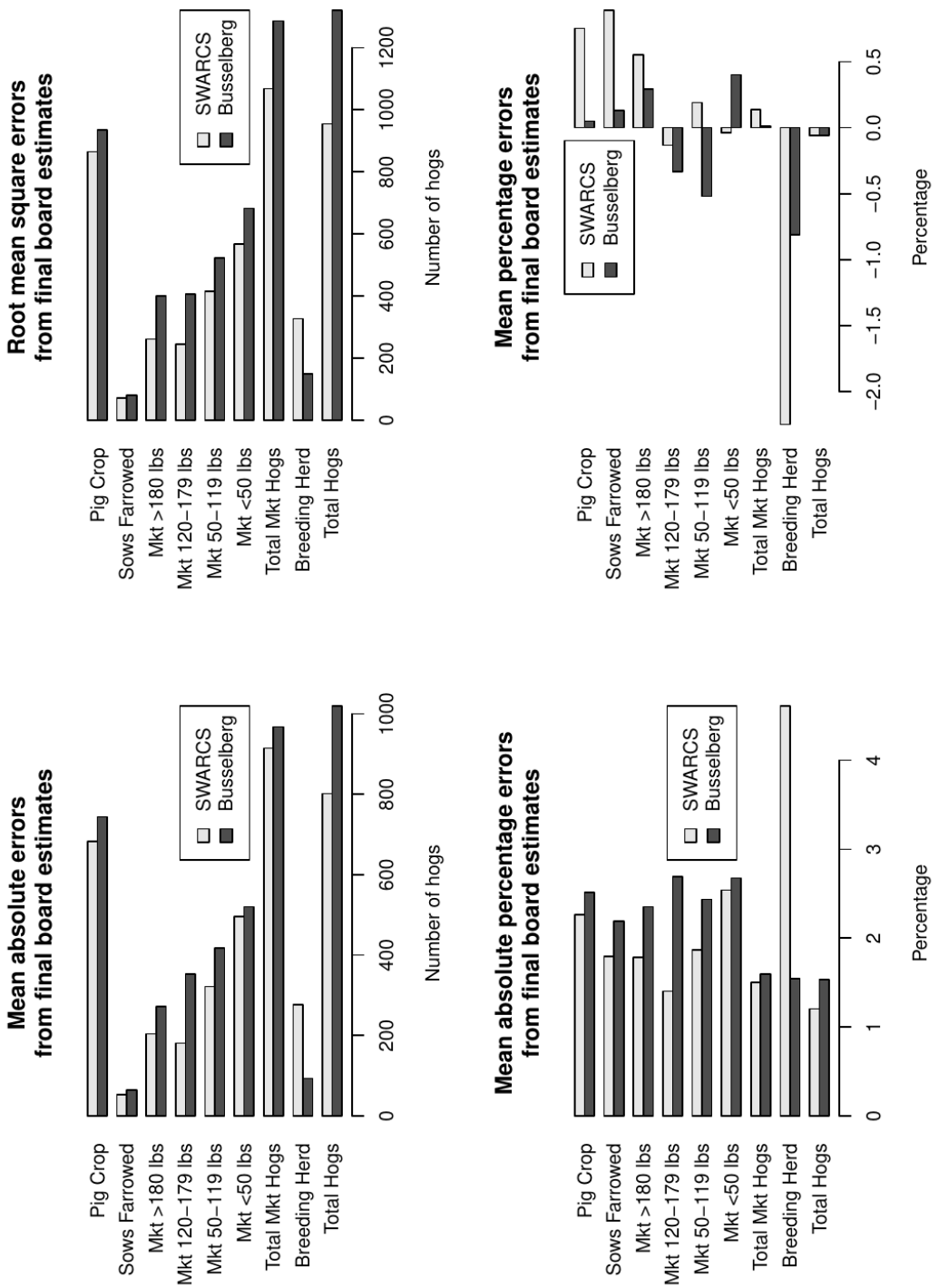

Figure 11: Prediction accuracy of the models when compared to the final board estimates.

7 Work Cited

Box, G.E.P., G.M. Jenkins, G.C. Reinsel, and G.M. Ljung. 2015. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.

Busselberg, S. 2013. "The Use of Signal Filtering for Hog Inventory Estimation." *Proceedings of the Federal Committee on Statistical Methodology (FCSM) Research Conference*.

Byrd, R., P. Lu, J. Nocedal, and C. Zhu. 1995. "A Limited Memory Algorithm for Bound Constrained Optimization." *SIAM Journal on Scientific Computing* 16 (5): 1190–1208. https://doi.org/10.1137/0916069.

Eilers, P.H.C., and B.D. Marx. 1996. "Flexible Smoothing with B-Splines and Penalties." *Statistical Science*. JSTOR, 89–102.

Fletcher, R. 1987. *Practical Methods of Optimization*. Wiley-Interscience Publication. Wiley.

Hyndman, R.J., and A.B. Koehler. 2006. "Another Look at Measures of Forecast Accuracy." *International Journal of Forecasting* 22 (4). Elsevier: 679–88.

Kedem, B., and L. Pan. 2015. *Time Series Prediction of Hog Inventory*. Unpublished internal document, United States Department of Agriculture NASS.

Khan, A., and W.B. Hildreth. 2003. *Case Studies in Public Budgeting and Financial Management*. Marcel Dekker Nova York.

NASS. 2005. *Estimation Manual Volume 4: Livestock and Dairy*. Unpublished internal document, United States Department of Agriculture NASS.

Pollard, J.H. 1966. "On the Use of the Direct Matrix Product in Analysing Certain Stochastic Population Models." *Biometrika* 53 (3-4). Oxford University Press: 397–415.

Roberts, Steven, and Gen Nowak. 2014. "Stabilizing the Lasso Against Cross-Validation Variability." *Computational Statistics & Data Analysis* 70. Elsevier: 198–211.

Rumelhart, D.E., G.E. Hinton, and R.J. Williams. 1985. "Learning Internal Representations by Error Propagation." California University San Diego, La Jolla Institute for Cognitive Science.

Shull, C.M. 2013. "Modeling Growth of Pigs Reared to Heavy Weights." PhD thesis, University of Illinois at Urbana-Champaign.

Tibshirani, Robert. 1996. “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1). Wiley Online Library: 267–88.

Wang, Hansheng, Guodong Li, and Chih-Ling Tsai. 2007. “Regression Coefficient and Autoregressive Order Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69 (1). Wiley Online Library: 63–78.

Chapter 6: Next? Options and Open Questions

Nell Sedransk

1 Today as the Starting Point

In the Overview the critical problems were identified as formulating a model or models that could encompass historically based patterns and basic hog biology in such a way that estimates of hog inventory would be coherent across time and would align with an external "gold standard."

The model system, whether a single model or multiple models that could be linked in some fashion, needs to be sensitive to change or shock without the lag of one or two quarters that is currently observed in both model estimates and ASB official estimates (with ASB revisions following after one or more quarters).

Currently, state estimates are allocations of the official national estimates back to the individual states. This is done by the ASB with knowledge of the state recommendations and other information; currently there are no model-based state estimates.

As investment in modeling hog inventory proceeds forward, there are options and open questions at several levels that require astute choices.

1.1 High-level Options and Questions

The first modeling commitment must be to the fundamental model structure. This involves the combining of information/model components of least three kinds: long-term patterns (historical data), functional patterns (biology of hog growth and survival), short-term patterns (disruptions). However these are to be combined, the logical requirements for relationships across quarters need to be met. It also involves choice of primary scale: national (top-down with expansion to allow state-level estimates), state-level/finer scale (with aggregation to national level estimates).

At one end of the spectrum is a comprehensive model incorporating all three components. At the other end of the spectrum is a set of models that include a primary (equilibrium) model so that disparities between the primary and the other models can serve to create a series of diagnostics and/or estimates of the divergences.

Current models utilize aggregate data (as adjusted by algorithm or by experts). A very different alternative approach associates an indicator function with each hog-producing operation (sampled or not) so that a probability (based on key factors such as geography, contagion pattern, local economics) could be assigned to each possible model from a set (equilibrium, disease outbreak, disaster, etc.). Aggregation would follow.

Open questions: How best to put the parts together?
Which primary scale for model – unit/state/nation?
How best to ensure coherence across quarters?

1.2 Next-level Options and Questions

The inclusion of covariate information, (particularly, state, geography for climate, geography for dynamics of disturbance, operation size, possibly operation type) can be at the unit, first-level aggregation (state), or high-level aggregation (national) level.

Open questions: How best to introduce biologic relationships into the model system?
How best to introduce spatial relationship into the model system?
How best to introduce covariates?

1.3 Specific Questions and Possible Options

Meaningful calculation of uncertainty is not easily defined for a model system that incorporates sampling estimators with model-based dynamics. Estimators for the design-based stratified sampling plan do not measure the same thing as model-based variance estimators.

The lag in detecting the impact of a disturbance or shock is documented by the ASB corrections of estimates for earlier quarters; the tested models similarly show lags. One reason is the small and localized impact of onset (a decrease of 35,000 hogs in the North Carolina inventory would be important to the state but within the 50,000-hog tolerance for national inventory). A second reason is a general conservatism in deviating from the expectation of equilibrium (a phenomenon observed in completely different federal and other data series), especially in the absence of external confirmation.

In the future web-scraping may provide a solution both at the national level and at state and county levels. Web-scraping has the advantage that it is not confined to the time frame for data processing. Also it can be conducted on a within-state level to provide information about extent of penetration of an outbreak or of impact from a natural disaster. Further, it can be linked to maps at the county or higher level (counties are available for all operations.)

Robustness of a model and coherence of estimates across time are important in periods of disequilibrium as well as period of equilibrium. Vetting a model is challenging, and especially so when available detailed data include few instances of national – scale disturbances. In view of the investment required to develop a model or model system fully, evaluating model performance cannot wait until development is complete.

Open questions:

How best to define a meaningful uncertainty measure for a model system that combines dissimilar components?

How best to detect occurrence and to estimate extent of disturbance impact on a finer scale?

How to vet the model from an early stage of development through completion, particularly to avoid overuse of the same testing framework or the same data base?

Of course additional questions will continue to arise throughout the process of model development. However the important open questions at this point are those that can set the direction for the modeling work to take.