


# Modeling swine population dynamics at a finer temporal resolution

Luca Sartore<sup>1,2</sup>  | Yijun Wei<sup>1,2</sup> | Emilola Abayomi<sup>2</sup> | Seth Riggins<sup>2</sup> | Gavin Corral<sup>2</sup> | Valbona Bejleri<sup>2</sup> | Clifford Spiegelman<sup>2,3</sup>

<sup>1</sup>National Institute of Statistical Science, Washington, DC, USA

<sup>2</sup>National Agriculture Statistics Service, United States Department of Agriculture, Washington, DC, USA

<sup>3</sup>Department of Statistics, Texas A&M University, College Station, Texas, USA

## Correspondence

Luca Sartore, National Institute of Statistical Science, 1400 Independence Ave. SW, Washington, DC 20250, USA.  
Email: lsartore@niss.org

## Abstract

The United States Department of Agriculture's National Agricultural Statistics Service (NASS) uses probability surveys of hog owners to estimate quarterly hog inventories in the United States at the national and state levels. NASS also receives data from external sources. A panel of commodity experts forms the Agricultural Statistics Board (ASB). The ASB establishes the NASS official estimates for each quarter by taking into account survey estimates and other relevant sources of information that are available in numerical and non-numerical form. The aim of this article is to propose an estimation method of hog inventories by combining the NASS proprietary survey results, the hog transaction data, the past ASB panel expert analyses, biological dynamics, and the inter-inventory relationship constraints. This approach downscales the official estimates to provide monthly estimates according to well-defined biological growth patterns. The model developed in this study provides national estimates that may inform the quarterly reports.

## KEYWORDS

hog inventories, modeling livestock, monthly swine estimates, population dynamics

## 1 | INTRODUCTION

National hog inventories constitute one of six principle economic indicators for the status of the United States (US) agricultural economy. The United States Department of Agriculture (USDA) National Agriculture Statistics Service (NASS) uses quarterly surveys to quantify hog inventories in the US at the national and state levels. The survey sampling approach is used to investigate the US swine population, establish unbiased estimates of inventories and track the evolution of the swine industry over time. However, survey estimates are not always consistent with the information available from other sources, such as biological growth, administrative data, and historical records. To address this issue, a panel of commodity experts forming the Agricultural Statistics Board (ASB) sets current official estimates by combining survey estimates, administrative data, and historical records.

Modeling the temporal dynamics of the swine population is a natural statistical avenue for providing timely and accurate inventory estimates with measures of uncertainty to the ASB. Model-based estimates are required to be consistent with hog biology and standard practices of the swine industry. A comprehensive formulation of the quantities to model that combines historical estimates with current survey data is essential for capturing temporal relationships across variables.

This article is structured as follows. Section 2 describes the current model, other attempts to model hog inventories, and the objectives addressed by the new model. Section 3 provides an overview of the available information and the data requirements to properly fit the model. The estimation process to produce national estimates for the swine report is presented in Section 4. The proposed model is introduced in Section 5. The estimation procedures adopted for producing model-based estimates are presented in Section 6. Data analyses and model comparisons are presented in Section 7. Final remarks and conclusions are given in Section 8.

## 2 | THE NEED FOR A NEW MODEL

New models can be developed to provide quarterly estimates. To facilitate this process, the new model should consider all available information within a comprehensive mathematical formulation. In particular, the model should be able to produce estimates for the variables of interest by tracking:

- Hogs population dynamics from birth to slaughter;
- Number of deaths due to disease outbreaks and survival rates;
- Economic cycles of expansion (e.g., when pork producers recover the losses experienced from disease outbreaks) and contraction.

To achieve these goals, NASS has been using a model developed by Busselberg,<sup>1</sup> which is a time series approach based on a constrained state-space model (SSM). Given the present information, the SSM provides inventory estimates by maintaining stable accounting relationships and biological dynamics and satisfying a set of constraints.<sup>2</sup> This approach consists of two phases:

1. The model parameters describing linear dynamics of the system and its evolution are estimated such that the trajectory of the model satisfies the given constraints;
2. The inventory estimates are based on the most current parameter estimates. Measures of uncertainty, such as standard errors and coefficients of variation, are also provided.

Another model was developed by Kedem and Pan<sup>3</sup> to improve the SSM and implement a more flexible model that is able to capture departures from a state of equilibrium during unstable periods. This approach is based on a sequence of pairwise comparisons between linear time series models with exogenous variables, where the best model is selected for predicting the quantities of the quarter of interest. Useful economic covariates, such as hog and pork prices, are tested using spectral analysis, and the final model is selected according to its efficacy. At the end of the process, the winning model is selected to produce estimates, forecasts, and the measures of uncertainty.

These two models are capable of producing estimates with desirable characteristics, but the strengths of one are the weaknesses of the other and vice-versa. Although the SSM takes into consideration biological properties of hogs and captures an equilibrium dynamic that satisfies the accounting constraints, it is unable to adapt quickly to systematic shocks, such as disease outbreaks, resulting in heavily biased and unrealistic results. On the other hand, Kedem and Pan<sup>3</sup> model provides a very 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 that adapts quickly to shocks and also takes into consideration the biological growth of the hogs forming the population under study is desired. The approach considered in Section 5 tracks the newborn piglets by both modeling their growth and survival rates. Other relationships, such as those between breeding herd and sows farrowed, are based on the knowledge of biological gestation of sows (e.g., durations and average litter sizes), and they are formulated separately.

### 2.1 | Finer temporal resolution

Hog reports are released to the public every quarter toward the end of March, June, September, and December. The official statistics show:

- The hog and pig inventory of the current quarter and the previous four;
- The inventory by class (i.e., breeding herd and total market hogs);
- The inventory of market hogs by weight group and by state;
- The number of sows farrowed, pigs per litter and pig crop in every quarter and month;
- The farrowing intentions by state for the following quarter.

The official report in December provides a detailed description of the quantities measured through the survey in all 50 states whereas the others focus on the 16 major hog producing states (specifically for Colorado, Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Carolina, Ohio, Oklahoma, Pennsylvania, South Dakota, Texas, and Utah). This framework allows NASS to quantify yearly coverage adjustments by more frequently sampling the 16 states that supply 95% of the total US swine production.

There is an on-going effort to improve the methodology used to compute model-based estimates. Survey data are collected on a quarterly basis, but questionnaire respondents enumerate sows farrowed and pig crop on a monthly basis. The other inventory variables are collected quarterly as point-in-time data. Thus, modeling monthly inventory numbers requires identifying the relationships between quarterly and monthly data and extrapolating monthly numbers when these are not directly surveyed.

From a biological perspective, the relationships between variables of interest are evident and self-explanatory. In particular, about a sixth of the sows in the breeding herd farrow every month. These sows produce a litter of 10 piglets on average. The new born litter will be then counted with all the other pigs weighing less than 50 lbs. As the hogs and pigs grow in age, they gain weight. USDA tracks quantities of hogs as they transition from one weight class to a heavier one. Moreover, it is also possible to include the average survival rate of hogs and introduce realistic variations that are expected to be dependent on the weight of the animals. This is reasonable since the younger hogs are more vulnerable to epidemics.

## 2.2 | Reactivity to shocks

Shocks impacting the number of estimated hogs are sudden departures from their expected values computed under standard conditions. These conditions are defined using an equilibrium pattern, which may reasonably extend over a long period of time. The occurrence and the impact of a shock are not easily predictable.

However, shocks are not the only issue with departures from the expected trajectories of the time series; in fact, abnormal cyclical patterns can manifest and become the new cyclical pattern for the future. The identification of systemic shocks or unforeseen changes of the process requires specific testing techniques, which can be combined with a flexible formulation of the model to improve the inventory estimation using available data.

Shocks are due mainly to either disease outbreaks or market phenomena. Elevated mortality rates are associated with specific diseases. Usually, estimated pig crop rapidly drops as the disease spreads, and increases to get back to normality afterwards. Some example of market phenomena are found in crises, new national and international trading policies connected to the pork demand, and other events that impact market prices, which may include climate-adverse conditions, variations of the capacity of the slaughter facilities, and new breeding techniques.

## 2.3 | Reducing number of constraints

The SSM,<sup>1</sup> as currently used at NASS, imposes constraints to produce reliable results that are consistent with hog biology and the available information at the national level. The limitation of this approach mainly consists of attaining maximum likelihood estimates on a restricted parametric space that is also shaped by the values assigned to the latent variables, which are unknown. This is quite difficult to achieve since the estimation of the model parameters is based on the standard Expectation-Maximization (EM) algorithm,<sup>4</sup> which iterates two steps until the usual convergence criteria are met. If this iterative procedure is performed without forcing constraints, the resulting values would contradict reasonable biological laws related to the herd reproduction, population growth, and survival. For example, it is possible to generate large numbers of hogs belonging in a heavy weight-category without having enough hogs in the lighter weight-groups during the previous months. Furthermore, if the model does not take into consideration physical, biological, and economic

constraints, contradictory totals are likely to appear either during the periods under the influence of shocks or after the epidemics when the effect of disease outbreaks is mitigated.

Through the inclusion of constraints, the algorithm produces results that are consistent with both the expected biological growth of hogs and other population dynamics related to birth and death processes. The constraints operate in two ways:

- Changing the direction and the magnitude of the descent step while optimizing the model parameters, for example, by introducing a penalty term when performing LASSO regressions;
- Forcing a set of linear and nonlinear equations defining known relationships among variables of the model, for example, the computation of the latent variables gets simplified by introducing constraints in the model of the outcome variables.

The combination of these two approaches allows a regression method to be developed that simultaneously produces estimates for the swine inventories and the model parameters that are coherent with the swine population dynamics. However, the SSM is computationally limited since the number of constraints causes the model to be very rigid. Furthermore, the EM algorithm converges slowly to a reliable solution especially when the constraints hardly admit a feasible one.

The mortality rate included in the SSM appears as an accounting constraint, and it remains fixed across time. Hence, this approach does not account for the risks of major diseases associated with hog casualties. Other accountability constraints were developed to avoid abnormal patterns during periods of long-term equilibrium, where both the observed data and past official statistics provide enough information to estimate inventory quantities. The mathematical constraints are forced during regression, resulting in parameter estimates that have overly rigid constraints and thus cannot change quickly when the system is in a state of disequilibrium, that is, when a shock changes the dynamics of the system.

Standard population dynamics have been explained in the past using simple stochastic models,<sup>5</sup> which can be further extended to address the rigidity of the SSM. Stochastic models also allow the complexity of the population dynamics to be studied by providing a distribution for the random components in the model.

### 3 | SOURCES OF INFORMATION

Survey data are collected through questionnaires and constitute the most important source of information. However, other sources such as administrative records and historical official statistics are also helpful in establishing temporal patterns, such as trend, seasonality or other cyclical behaviors that are stable across time. Based on all these sources of data, the ASB produces a set of reliable estimates that account for the modeled values, administrative data, and non-numerical knowledge. A new mathematical approach determines the population dynamics through the use of the past official estimates and makes adjustments that account for summary statistics from current survey data.

#### 3.1 | Published estimates

NASS publishes quarterly statistics for hogs and pigs. These reports are organized in sections and primarily consist of tables that show the official statistics for sows farrowing, pig crop, pigs-per-litter, breeding herd, market inventory by weight group and total inventory. These tables are provided for the entire nation and the 16 major hog producing states (or all states for the report of December). Monthly statistics are also provided only for sows farrowed, pig crop, and pigs-per-litter at the national level.

Quarterly swine inventory quantifies the total hogs and pigs in the US as of the first day of March, June, September, and December. These totals are then split into two quantities associated with two hog classes: the breeding herd and market hogs. The numbers of sows and boars kept for breeding are distinguished from the hogs raised to be marketed. The market hogs are themselves subdivided into the following weight groups:

- hogs weighing under 50 pounds;
- hogs weighing between 50 and 119 pounds;

- hogs weighing between 120 and 179 pounds;
- hogs weighing at least 180 pounds.

The number of sows farrowed and newborn piglets are provided over semiannual, quarterly, and monthly periods. Their estimates are published together with the litter rate computed as the ratio between sows farrowed and litter size. These quantities are useful in establishing production patterns at several temporal resolutions.

The ASB revises the past four quarters to ensure that the final estimates are consistent with the totals provided by administrative data. This process can be seen as a post-calibration to reduce the potential bias of the estimates already published, for example, the slaughter data are collected weekly and can be used to inform adjustments to the estimates.

### 3.2 | Survey data

Survey data are collected from a stratified random sample of hog and pig producers. Stratification is applied based on control data for the number of hogs owned by the operation. The stratified samples are selected according to sample-size allocations defined by state and strata of homogenous operators. The December Hog Survey is the only quarter survey for which the operations of all states can be sampled. NASS collects data by mail, computer assisted web interviewing (CAWI), computer assisted telephone interviewing (CATI), and computer assisted personal interviewing (CAPI). Phone follow-up of mail non-response is also conducted to reach the highest response rate.

The survey is designed to collect data for several variables that monitor the production of hogs and pigs in the US. The main variables used for producing quarterly/monthly estimates are

- the number of sows farrowed over the three consecutive months preceding the surveyed quarter (one measurement per month);
- the size of the litter (post-weaning) measured over the same periods for the sows farrowed;
- the size of the breeding herd (number of heads counted on the first day of the surveyed quarter). It mainly consists of sows, even if each operation keeps a negligible number of boars for breeding purposes. The sows in the breeding herd are either bred, weaning piglets, or between gestation periods, which last about three months;
- the number of the market hogs distinguished according to the four weight classes (number of heads counted on the first day of the surveyed quarter); and
- the breeding intention for the next two quarters.

The size of an operation plays a role in the frequency of data collection for specific operations and the completeness of data that are collected. These factors are unpredictable when a disruption of equilibrium occurs. Further, reported data are seldom complete so imputation techniques are utilized to account for the missing data.

### 3.3 | Biological specification

The biological patterns of growth have been thoroughly studied as exemplified by Shull,<sup>6</sup> Park et al.,<sup>7</sup> and Park and Oh.<sup>8</sup> An average growing pattern under relatively normal circumstances can establish the time required for hogs and pigs to reach a weight of 250 lbs, which is approximately the weight threshold that makes a hog ready to be processed by slaughter facilities.<sup>9</sup> From the analysis of the results in table 23 (page 117 of Shull<sup>6</sup>), the model proposed by Bridges et al.<sup>10</sup> best describes the average growth pattern of hogs and is mathematically

$$W(t) = \omega_0 + \omega_1(1 - \exp(-\omega_2 t^{\omega_3})),$$

where  $W(t)$  is the average weight of a living hog at a time  $t$  expressed in months from the date of birth,  $\omega_0$  represents the mean weight at weaning (approximately 12.57 lbs),  $\omega_1$  denotes the mean-growth upper asymptote (approximately 440.04 lbs),  $\omega_2$  and  $\omega_3$  are parameters that control the shape of the growth curve. The value of  $\omega_2$  represents the growing rate of approximately 0.0166, and  $\omega_3$  denotes the acceleration of the growth of approximately 1.9858.

The average biological growth determines the number of times a hog belonging to a certain weight class is counted during some time frame. This approach allows the number of hogs that transition from one weight class to another within a specific time frame to be quantified. These transitions are then used to calculate the propagation of the newborn piglets at several measurement days across time. In so doing, the estimates of the market hogs by weight group will be more inclined to satisfy the constraints imposed by the current SSM.

### 3.4 | Administrative sources

Slaughter data are the primary 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 of the regulated slaughter facilities. The number of pork carcasses are enumerated and can be combined with other datasets to enhance the analysis.

These data are available on a weekly basis and are part of the review of published estimates. The data 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.

These data are not used as predictors to develop the model, but they can be employed to constrain the heaviest inventory class of hogs. Instead of using the past official estimates in the model, the most recent values attributed to the variables involved in the analysis are adjusted within the model by aggregating the weekly slaughter numbers and the information provided by the official statistics. This offers the opportunity to adjust the trajectory of the model, under the assumption that the best use of the available information is made.

## 4 | OVERVIEW OF THE ESTIMATION PROCESS

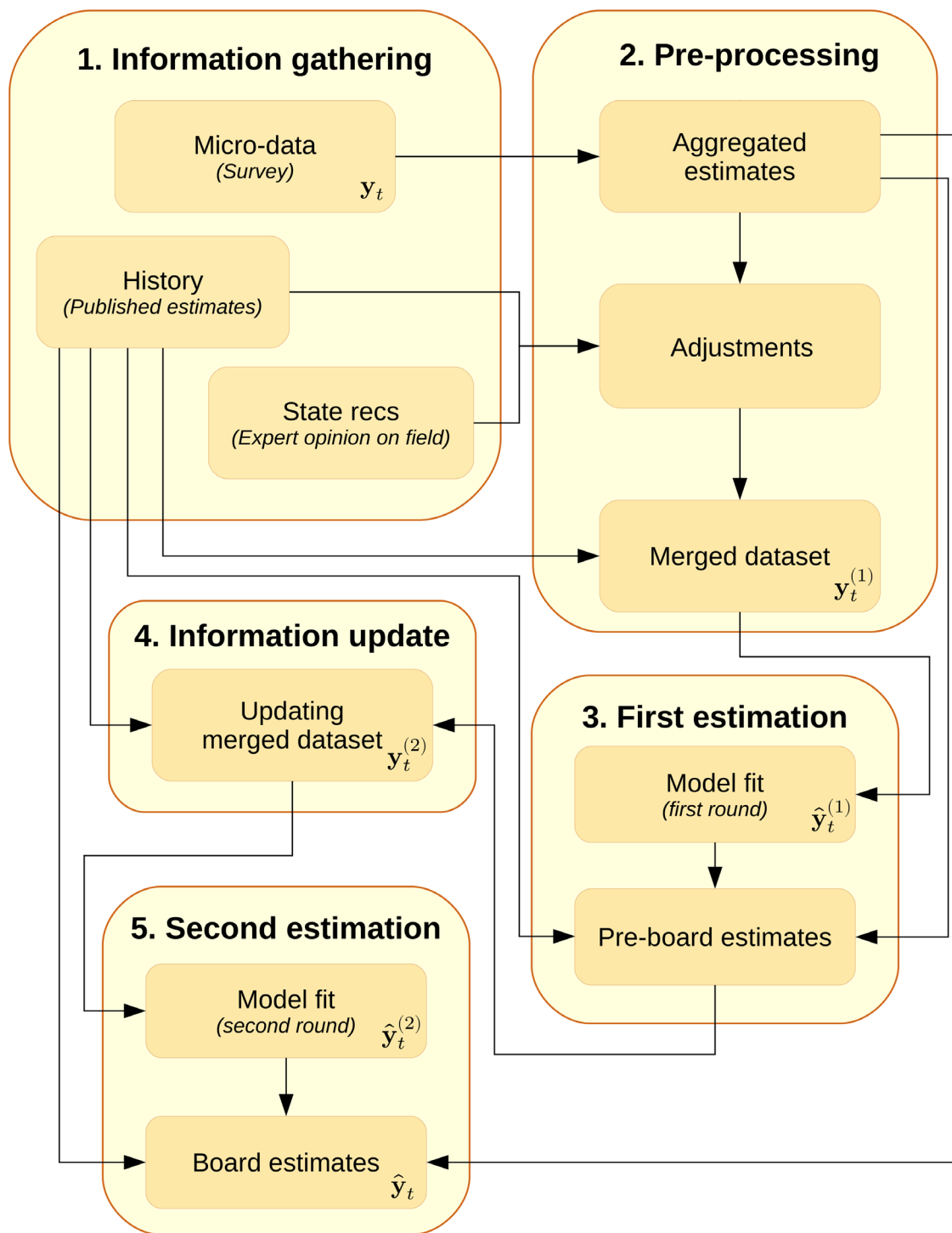
The proposed model provides monthly estimates of hog inventories at a national level by modeling biological dynamics for the US swine population. The estimation process of estimating  $\hat{y}_t$  consists of five stages (Figure 1):

1. *Information gathering*: organizing preliminary information;
2. *Pre-processing*: adjusting and summarizing the initially gathered information into a single dataset;
3. *First estimation*: producing estimates for the pre-board;
4. *Information update*: updating the dataset to be used after the pre-board;
5. *Second estimation*: producing updated estimates for the Agriculture Statistics Board (ABS).

The first stage consists of *gathering information* from the survey respondents at time  $t$  in the form of micro-data. NASS's official estimates are based on historical and administrative data and state recommendations provided by NASS's field offices. All this information is organized and made available for successive adjustments.

The *pre-processing* stage consists of three 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 current survey data. During this stage, micro-data are aggregated into summary statistics that are adjusted for coverage, non-response, and sampling errors.

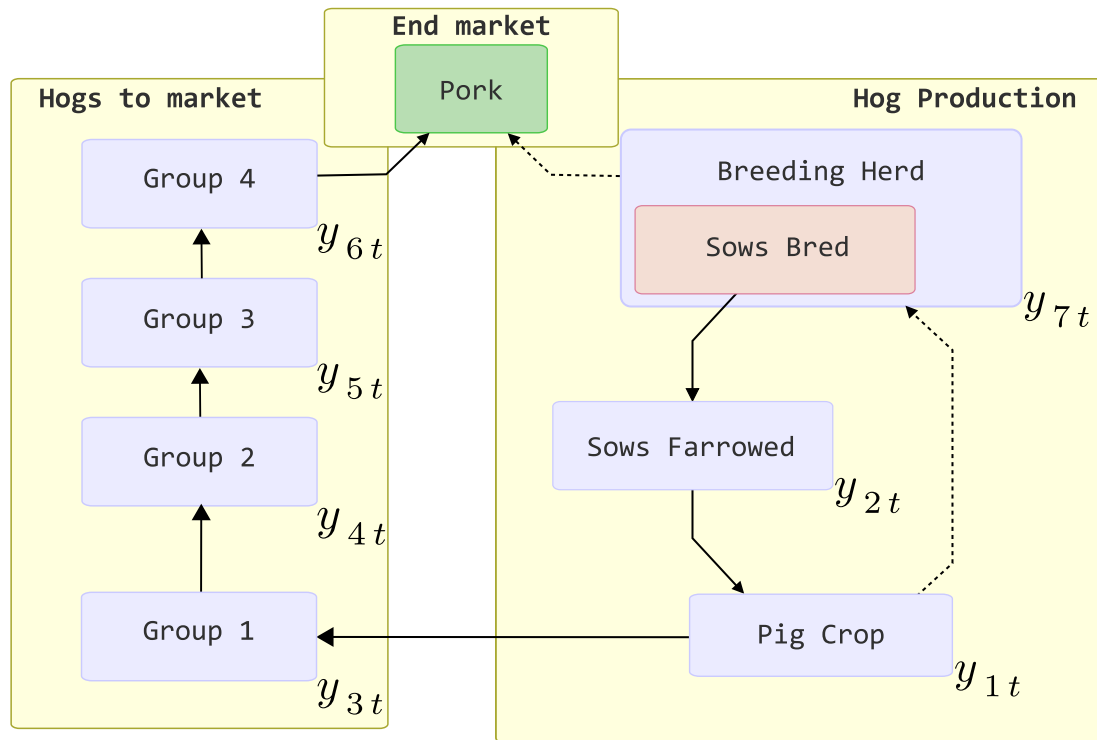
Once a comprehensive dataset  $\mathbf{y}_t^{(1)}$  is created, the *first estimation* process starts. The parameters of the model are estimated by using iterative regression techniques, and the fitted values for the variables of interest  $\hat{y}_t^{(1)}$  are calculated for the most recent quarter.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



**FIGURE 1** Estimation process [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

The pre-board provides a set of estimates  $y_t^{(2)}$  for the current quarter and revises values of the published estimates in the last four quarters based on administrative data. These can inform the estimates by providing updates that account for information acquired one year after the publication date of first estimates. The published statistics can be considered final after the fourth revision.

Once the dataset is fully updated, 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  $\hat{y}_t^{(2)}$  are provided to the ASB. The ASB consists of nine or ten livestock-commodity experts (including those forming the pre-board) who set the official estimates  $\hat{y}_t$ .



**FIGURE 2** Pork production processes. The solid arrows represent the dynamics considered in the SDTP model. The dashed arrows denote existing dynamics that do not have the potential to alter the results when ignored [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

## 5 | A STRUCTURED DISCRETE TIME POPULATION MODEL

A structured discrete time population (SDTP) model<sup>11</sup> can track the growth of newborn piglets, and provide monthly estimates for the inventory number of market hogs (classified by weight). In an approach different from the Liz and Pilarczyk's proposal,<sup>11</sup> the survival rates associated with monthly cohorts of newborn piglets and harvest rates are used only for swine that reached proper maturity weight. Also, the stock-recruitment function is handled differently, since the size of the breeding herd, pig crop (i.e., the number of weaned piglets), and the number of sows farrowed are determined with classical time-series models. The SDTP model presented in this section assumes an average dynamic growth rate for weaned pigs born within a month, and it reproduces patterns of 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 that managers use to establish and improve production efficiency. This analysis leads to a simple model that describes the connections among the variables of interest (see Figure 2) and honors biological constraints.

The model is divided into two systems 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 and provide monthly estimates that can be used to track hog production on a finer time resolution.
- The second describes 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. However, the size of the breeding herd is only part of the first system due to the close relationship with the number of sows farrowed (see Figure 2).

The proposed SDTP model will be explained in the following sections.



## 5.1 | Model for variables enumerated monthly

At the national level, pig crop, sows farrowed, and litter rate are modeled differently than the five basic inventory items. The strategy of having two separate models allows for different time units (monthly for pig crop and sows farrowed versus quarterly for the other variables) and provides a reasonable explanation of the hog population dynamics from a macroscopic perspective.

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

$$\begin{cases} E[y_{1,t}] = \rho_t E[y_{2,t}], \\ y_{2,t} = \varphi y_{7,t-2} + \varepsilon_{2,t}, \end{cases} \quad (1)$$

where time  $t$  is expressed in months,  $y_{1,t}$  denotes monthly pig-crop,  $y_{2,t}$  represents monthly sows farrowed,  $\rho_t$  indicates the litter rate at time  $t$  as it appears in NASS report,  $y_{7,t-2}$  expresses the breeding herd size as measured on the first day of the month  $t-2$ ,  $\varphi$  denotes the farrowing rate, and  $\varepsilon_{2,t}$  accounts for the statistical error in modeling monthly sows farrowed (as represented in Figure 2). However, since the values of  $y_{7,t}$  and  $y_{2,t}$  are, respectively, available on a quarterly and monthly basis, the estimates for the breeding herd are obtained as  $y_{7,t} = \varphi^{-1} y_{2,t+2} + \varepsilon_{7,t}$ , where  $\varepsilon_{7,t}$  denotes the monthly errors of the breeding herd.

The dynamics of  $\log(y_{1,t})$  and  $\log(y_{2,t})$  are both modeled by a Seasonal AutoRegressive Integrated Moving Average (SARIMA) model.<sup>12</sup> In particular, a SARIMA(2, 1, 2)  $\times$  (2, 1, 2)<sub>12</sub> is fit using LASSO regression.<sup>13</sup> 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 having 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 et al.,<sup>14</sup> also allows for automatic time series model selection to be used in the estimation of the logarithms of pig crop,  $\log(y_{1,t})$ , and sows farrowed,  $\log(y_{2,t})$ . Thus, in addition to Equation (1), the following set of equations should be considered in the estimation process:

$$\begin{cases} (1 + \phi_{1,1}B + \phi_{1,2}B^2)(1 + \phi_{1,12}B^{12} + \phi_{1,24}B^{24})\nabla\nabla_{12} \log(y_{1,t}) = (1 + \theta_{1,1}B + \theta_{1,2}B^2)(1 + \theta_{1,12}B^{12} + \theta_{1,24}B^{24})\tilde{\varepsilon}_{1,t}, \\ (1 + \phi_{2,1}B + \phi_{2,2}B^2)(1 + \phi_{2,12}B^{12} + \phi_{2,24}B^{24})\nabla\nabla_{12} \log(y_{2,t}) = (1 + \theta_{2,1}B + \theta_{2,2}B^2)(1 + \theta_{2,12}B^{12} + \theta_{2,24}B^{24})\tilde{\varepsilon}_{2,t}, \\ \log(\rho_t) = \frac{(1 + \theta_{1,1}B + \theta_{1,2}B^2)(1 + \theta_{1,12}B^{12} + \theta_{1,24}B^{24})\tilde{\varepsilon}_{1,t}}{(1 + \phi_{1,1}B + \phi_{1,2}B^2)(1 + \phi_{1,12}B^{12} + \phi_{1,24}B^{24})\nabla\nabla_{12}} - \frac{(1 + \theta_{2,1}B + \theta_{2,2}B^2)(1 + \theta_{2,12}B^{12} + \theta_{2,24}B^{24})\tilde{\varepsilon}_{2,t}}{(1 + \phi_{2,1}B + \phi_{2,2}B^2)(1 + \phi_{2,12}B^{12} + \phi_{2,24}B^{24})\nabla\nabla_{12}}, \end{cases} \quad (2)$$

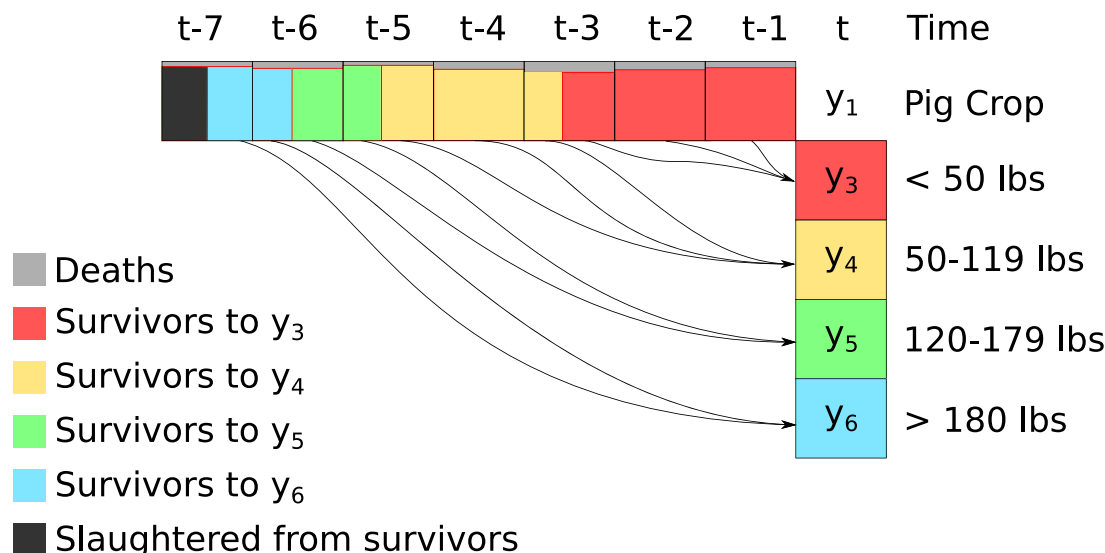
where  $B^h$  denotes the backward operator of  $h$  steps (e.g., the notation  $B^3 y_{1,t}$  is equivalent to  $y_{1,t-3}$ ),  $\nabla_S^d = (1 - B^S)^d$  corresponds to the difference operator of order  $d$  at lag  $S=12$  (e.g., the notation  $\nabla_{12}^3 \zeta_t = (1 - B^{12})^3 \zeta_t$  is equivalent to  $\zeta_t - 3 \zeta_{t-12} + 3 \zeta_{t-24} - \zeta_{t-36}$ ). Since  $d$  and  $S$  act as powers in a polynomial involving the backward operator  $B$ , they can be omitted when their value is one (for additional details, please, see pages 308 and 310 in Box et al.,<sup>12</sup> or Box and Jenkins<sup>15</sup>). The parameters  $\theta_{i,j}$  and  $\phi_{i,j}$  are, respectively, associated with the effects of the auto-regressive and moving-average components of the variable  $i \in \{1, 2\}$  at the temporal lag  $j > 0$ . The residuals  $\tilde{\varepsilon}_{i,t}$  correspond to the error in predicting the variable  $i$  at time  $t$ .

## 5.2 | Model for monthly inventories

Similar to the proposal of Pollard,<sup>16</sup> the equations governing the behavior of the weight classes are defined as:

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

with the cohort allocation parameter  $\alpha_i \in [0, 1]$ , for any  $i = 1, \dots, 4$ ; the survival rate  $\zeta_t \in [0, 1]$  is associated with the monthly cohort  $y_{1,t}$ , such that the adjusted values of pig crop are propagated by accounting for pig losses within each



**FIGURE 3** Example of hog growth dynamics of the US swine population [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

cohort (see Figure 3); while  $y_{3,t}$  denotes the number hogs weighed less than 50 lbs on the first day of the month  $t$ , and similarly,  $y_{4,t}$  is used for hogs between 50 and 119 lbs,  $y_{5,t}$  for hogs between 120 and 179 lbs, and  $y_{6,t}$  for hogs weighing at least 180 lbs.

The cohort allocation parameter  $\alpha_1$  is used to split the weaned piglets at time  $t-3$  since  $W^{-1}(50) \approx 2.33$ , that is, the inverse function of the Bridges' model<sup>10</sup> presented in Section 3.3 computed for the upper bound of the first weight class. This means that only a fraction of the piglets born three months earlier moves into the next weight class. Similarly,  $\alpha_2$  splits the cohort born at time  $t-5$ , since  $W^{-1}(120) \approx 4.15$ ;  $\alpha_3$  splits the cohort born at time  $t-6$ , since  $W^{-1}(180) \approx 5.44$ ; and  $\alpha_4$  splits the cohort born at time  $t-7$ , since  $W^{-1}(250) \approx 6.93$ .

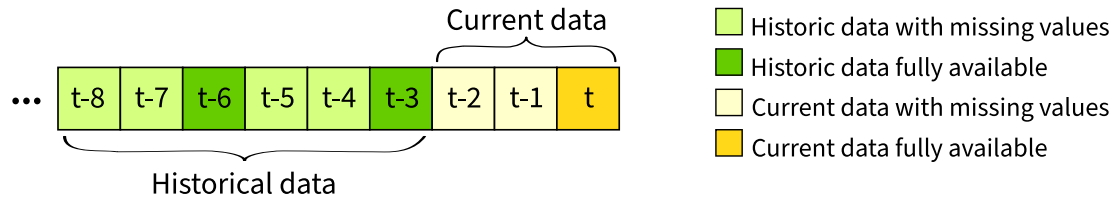
The relationships in (3) constrain the number of hogs in each weight group to be consistent with the number of piglets born in the past that are still alive. This formulation characterizes the survival probabilities of each monthly cohort during its lifespan. The simplified system of equations (3) can be extended by considering additional effects from lagged residuals, and/or including nonlinear terms. To avoid an over-parameterization of the model, the contribution of additional terms is not considered here.

The SDTP model (3) allows for a flexible formulation that can track lasting changes in monthly pig cohorts. The survival rates are cohort dependent and they are restricted also by minimizing the absolute values of the lagged differences  $\nabla^d \zeta_t$ , for  $d=1, \dots, 3$ . This approach has been inspired by the use of penalties as formulated in the P-spline proposal of Eilers and Marx<sup>17</sup> to maintain simple models without over-fitting the data. This technique can provide smooth survival rates that quickly adapt by accounting for the temporal evolution of the US swine population. For example, a cohort of piglets born during month  $t$  has a survival rate  $\zeta_t$  that is localized in time, and quantifies the chances of being alive up to the moment the cohort enters the slaughter facility. Low values will be obtained for epidemic periods. Typical survival rates have been estimated at the national level to about 95%.

## 6 | ESTIMATION PROCEDURE

The estimation of the model parameters occurs in two stages (see Figure 1). The aim of the first stage is to produce initial results by combining historical dynamics, the survey data, and the state recommendations. This is achieved by estimating the model parameters on a customized dataset (see Figure 4) consisting of historical data and information from the current quarter.

Current survey data at the record level are aggregated at the national level by accounting for the stratified sampling design, non-response, and state recommendations. Micro data  $y_{k,t,i}$ , for each variable  $k=1, \dots, 7$ , are used in a weighted sum



**FIGURE 4** Representation of the information flow in the dataset used for regression [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

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

where  $n_t$  denotes the sample size at time  $t$ , the factor  $r_{k,t}$  represents a calibrated ratio adjustment that accounts for the state recommendations, and the survey weights  $w_{t,i}$  are associated with the adjustments accounting for the incompleteness of the list frame, the sampling probabilities, and the lack of response from some sample units.<sup>18</sup> In particular, the survey weight can be decomposed as

$$w_{t,i} = \frac{N_{j,t}}{n_{j,t}} v_{j,t} \frac{n_{j,t}}{a_{j,t}}, \quad (4)$$

for any sample unit  $i$  in stratum  $j$ . The first factor in (4) corresponds to the inverse of the sample inclusion probability computed as the ratio between  $n_{j,t}$  (the sample size from stratum  $j$  at time  $t$ ) and  $N_{j,t}$  (the size of the stratum  $j$  at time  $t$ ); the scalar  $v_{j,t}$  denotes the inverse coverage probability of stratum  $j$  at time  $t$  to account for the records that are not on the NASS list frame; the last factor represents the inverse of the response probability computed as the ratio between  $a_{j,t}$  (responding sample units from stratum  $j$  at time  $t$ ) and  $n_{j,t}$ .

In the first estimation stage, the SDTP model uses the survey summaries,  $\mathbf{y}_t^{(1)}$ , to compute initial fitted values for the variables of interest,  $\hat{\mathbf{y}}_t^{(1)}$ . These outcomes are then evaluated by the pre-board, which produces a set of preliminary results,  $\mathbf{y}_t^{(2)}$ . Afterward, the SDTP model uses the values provided by the pre-board,  $\mathbf{y}_t^{(2)}$ , as the most reliable source of information for the three months in the current quarter (see Figure 4). The fitted values computed during the second estimation stage,  $\hat{\mathbf{y}}_t^{(2)}$ , are then used by the ASB to produce the official statistics,  $\hat{\mathbf{y}}_t$ .

The proposed time series methodology consists of two algorithms that, respectively, produce estimates for inventory items (i.e., for variable  $y_{k,t}$ , where  $k = 3, \dots, 7$ ) and non-inventory items (i.e., for variable  $y_{1,t}$  and  $y_{2,t}$ ). Both algorithms are iterative in nature and take advantage of known methods for solving nonlinear optimization problems. Thus, one starts with an initial guess for each of the parameter values that is updated by adding or subtracting non-negative quantities computed along a descending direction. These adjustments produce values with a smaller sum of squared residuals. From this new set of values another descending direction is computed. The process is repeated several times, until no further adjustments are required to reduce the model error. The iterative procedure uses the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm<sup>19</sup> to optimize the parameters and update the residuals of the equations in (2). On the other hand, the limited memory algorithm for bound constrained optimization<sup>20</sup> is applied to optimize the parameters in the system of equations (3). A simulation study is presented in Appendix A to demonstrate the stability of the estimation processes presented in Sections 6.1 and 6.2.

## 6.1 | Optimization for non-inventory items

To optimize the total loss associated with estimation of the parameters in (2), the quantity

$$\sum_{t=1}^{t_0} \left[ \sum_{k=1}^2 (\tilde{\varepsilon}_{k,t})^2 + \gamma (\tilde{\varepsilon}_{\rho,t})^2 + \sum_{\ell=0}^2 (y_{2,t-3\ell+2} - \varphi y_{7,t-3\ell})^2 \right] + \delta \sum_{k=1}^2 \sum_{i=0}^1 \sum_{j=1}^2 (|\phi_{k,12^j}| + |\theta_{k,12^j}|), \quad (5)$$

is minimized with respect to the parameters  $\phi_{k,12^j}$  and  $\theta_{k,12^j}$ , for  $k = 1, 2$ , and  $i = 0, 1$ , and  $j = 1, 2$ . These parameters have an impact on the residuals  $\tilde{\epsilon}_{k,t}$ , for  $k = 1, 2$ , and  $\tilde{\epsilon}_{\rho,t}$ , which represents the residual of the litter rate at time  $t$  in logarithmic scale, that is,  $\tilde{\epsilon}_{\rho,t} = \tilde{\epsilon}_{1,t} + \tilde{\epsilon}_{2,t}$ . The non-negative scalar  $\gamma$  governs the importance of the error associated to the litter rate, and it can be established on historical data by applying standard cross-validation methods.<sup>21</sup> In particular, one already controls for the error  $\tilde{\epsilon}_{\rho,t}$  by minimizing the sum of the square residuals governing the pig crop and the sows farrowed. However, better estimates can be produced for  $\gamma > 0$ , which is kept fixed over each regression fit to reduce the amount of computations. The penalty  $\delta$  is used when performing LASSO regression<sup>13</sup> on the time series models. By minimizing (5), LASSO regression is simultaneously performed on the three equations in (2), and it also accounts for the expressions of the system (1). This makes it possible to simultaneously estimate the parameters by accounting for inter-relationships that affect the behavior of other variables.

The solution of the problem stated in (5) is obtained through iterative algorithms, which require a stable set of initial values to start the minimization. The initial choice of values at the start of the estimation algorithm is  $\varphi = \frac{1}{6}$ , an approximation of the proportion of sows farrowed in a month from the breeding herd. The time series parameters are set  $\phi_{k,12^j} = 0$ , and  $\theta_{k,12^j} = 0$ , for  $k = 1, 2$ , and  $i = 0, 1$ , and  $j = 1, 2$ , reflecting the equilibrium of a static process. The initial values of the residuals  $\tilde{\epsilon}_{k,t}$  and  $\tilde{\epsilon}_{\rho,t}$  are also set to zero, and updated at each iteration. These choices have deep consequences on the model selection and the convergence of the algorithm. In fact, a different set of values can produce sub-optimal results or may induce the numerical algorithm to prematurely converge to local minimum.

The optimization of the quantity in (5) is conducted for each value of  $\delta$  in the set  $\Delta = \{0.8^i : i = 0, \dots, 40\}$  by performing the following steps:

1. 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 (5) becomes smaller;
2. Given the new values of the parameter, produce new values of residuals;
3. Determine whether the convergence is achieved. If not, repeat step 1 and 2 until convergence.

The chosen BFGS algorithm produces adjustments along a descent direction by efficiently approximating the curvature of the quantities to minimize. This approach provides a computationally efficient approximation to the closed-form solutions of a quadratic representation of the quantity to minimize, which provides a descent direction. The quadratic representation is based on the second order approximation provided by the Taylor series expansion of the quantity to minimize. It has been shown that the approximated solution of the quadratic form produced at each iteration converges to the optimal point.<sup>19</sup>

Once parameter estimates are produced for the specified values of the penalty  $\delta$ , the 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. Non-significant parameters are chosen by a voting system. Each parameter in the SARIMA model is estimated 41 times (accordingly to the cardinality of the set  $\Delta$ ). A binary vote, corresponding to each parameter estimate, is determined by testing whether the estimate is significantly different from zero. The parameters that are forced to be zero have a number of votes that does not match the highest number of votes among all the parameters.

This algorithm is also used to process the data to be used in the second estimation stage. As explained in Section 4, the new values of historical and adjusted statistics after the pre-board sets their updated values for the current and past four quarters are processed in the second estimation stage.

## 6.2 | Optimization for inventory items

To reduce the sum of squared residuals associated with estimation of the parameters in (3), the following quantity is minimized:

$$\frac{1}{t_0} \sum_{t=1}^{t_0} \sum_{k=3}^6 (\epsilon_{k,t})^2 + \psi \sum_{t=1}^{t_0} \left( |\zeta_t - 0.95| + \sum_{d=1}^3 |\nabla^d \zeta_t| \right), \quad (6)$$

such that  $\alpha_i, \zeta_t \in [0, 1]$ , for any  $i = 1, \dots, 4$  and  $t \in \mathbb{Z}$ . All parameters  $\alpha_i$  and  $\zeta_t$  govern the behavior of the residuals  $\varepsilon_{k,t}$  as defined in (3), for  $k = 3, \dots, 6$ . The non-negative scalar  $\psi$  is used to control for the size of the penalties. Guntuboyina et al.<sup>22</sup> provided theoretical results and optimality conditions of the estimator obtained by minimizing (6).

As for the monthly estimates, the choice of initial values to start the optimization procedure has important implications on the convergence of the algorithm, and the optimality of the results. For the equations in (3), the initial choice of the parameters  $\alpha_i$  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,<sup>6</sup> 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  $\zeta_t$  are set to 0.95, for any  $t \in \mathbb{Z}$ , so as to represent a hypothetical case 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.<sup>20</sup> allows for simultaneous minimization of the quantity (6) with respect to all parameters involved in the system of equations (3). This approach guarantees that the final results satisfy the boundary constraints set by the model. Under the assumption that the dataset used for regression reflects the true status of the swine population, the proposed methodology can quickly adapt to shocks and produce more reliable results (see Section 7 for the performance evaluation of this model).

### 6.3 | Variances of the response variables

The variances of the variables of interest are computed using the delta method:<sup>23</sup>

$$\text{VAR}[\hat{\mathbf{y}}_t] \approx \mathbf{G} \text{VAR}[\hat{\boldsymbol{\theta}}] \mathbf{G}^\top,$$

where the Jacobian matrix  $\mathbf{G}$  consists of the partial derivatives of the fitted model  $h(\cdot, \cdot)$ , that is,  $g_{ij} = \frac{\partial}{\partial \theta_j} h(\boldsymbol{\theta}, \mathbf{x}_i)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ , for any parameter  $j = 1, \dots, p$ , and set of covariates  $i = 1, \dots, n$ . The covariance matrix of the estimated parameters  $\text{VAR}[\hat{\boldsymbol{\theta}}]$  is computed through a low-rank approximation of its spectral decomposition. This approach is similar to the proposal of Fan et al.,<sup>24</sup> who derived the robust properties of the covariance matrix estimator. Cape et al.<sup>25</sup> discussed the asymptotic properties of low-rank approximations showing the unbiasedness and normality of the limiting distribution. Thus, the matrix  $\text{VAR}[\hat{\boldsymbol{\theta}}]$  is computed by using the spectral decomposition of the positive semidefinite Hessian matrix  $H(\hat{\boldsymbol{\theta}}) = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top$ , such that

$$\text{VAR}[\hat{\boldsymbol{\theta}}] \approx \mathbf{V}_* \boldsymbol{\Lambda}_*^{-1} \mathbf{V}_*^\top,$$

where  $\mathbf{V}$  is the matrix of the eigenvectors of Hessian, and the matrix  $\mathbf{V}_*$  denotes a sub-matrix of eigenvectors associated with the positive eigenvalues of  $H(\hat{\boldsymbol{\theta}})$ . The notation  $\boldsymbol{\Lambda}$  denotes the diagonal matrix of the eigenvalues of the Hessian matrix,  $H(\hat{\boldsymbol{\theta}})$ , while the diagonal matrix  $\boldsymbol{\Lambda}_*^{-1}$  has its diagonal entries equal to the inverse of the positive eigenvalues of  $H(\hat{\boldsymbol{\theta}})$ .

## 7 | DATA ANALYSES

The hog survey is conducted every year in March, June, September, and December. The reference date for the survey is the first day of the survey month. NASS uses a dual frame approach, consisting of the Hog Survey list frame 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 known 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 approximately 3% undercoverage of the list frame. 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. 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. The constrained SSM<sup>1</sup> is currently used at NASS to produce quarterly estimates, and it is relatively inflexible and fails when shocks occur. In fact,

the input data cannot override the rigid constraints nor a fixed survival rate. This results in a lag of at least one quarter in detecting a shock.

The proposed SDTP model is compared to both the SSM and Kedem and Pan's model<sup>3</sup> using classical model selection criteria. In particular, Hyndman and Koehler<sup>26</sup> provided a detailed review about measures of accuracy. In general, the criteria adopted to assess the performance of a regression model include (but are not restricted to) the following measures:<sup>26,27</sup>

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

$$\text{MAE}_k = \frac{1}{T} \sum_{t=1}^T |y_{k,t}^* - \hat{y}_{k,t}|. \quad (7)$$

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:

$$\text{RMSE}_k = \sqrt{\frac{1}{T} \sum_{t=1}^T (y_{k,t}^* - \hat{y}_{k,t})^2}. \quad (8)$$

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:

$$\text{MAPE}_k = \frac{100\%}{T} \sum_{t=1}^T \left| \frac{y_{k,t}^* - \hat{y}_{k,t}}{y_{k,t}^*} \right|. \quad (9)$$

This index reports the relative distance between fitted and true values as a percentage. MAPE is also robust to outliers as MAE.

- *Mean percentage error* (MPE) is computed as:

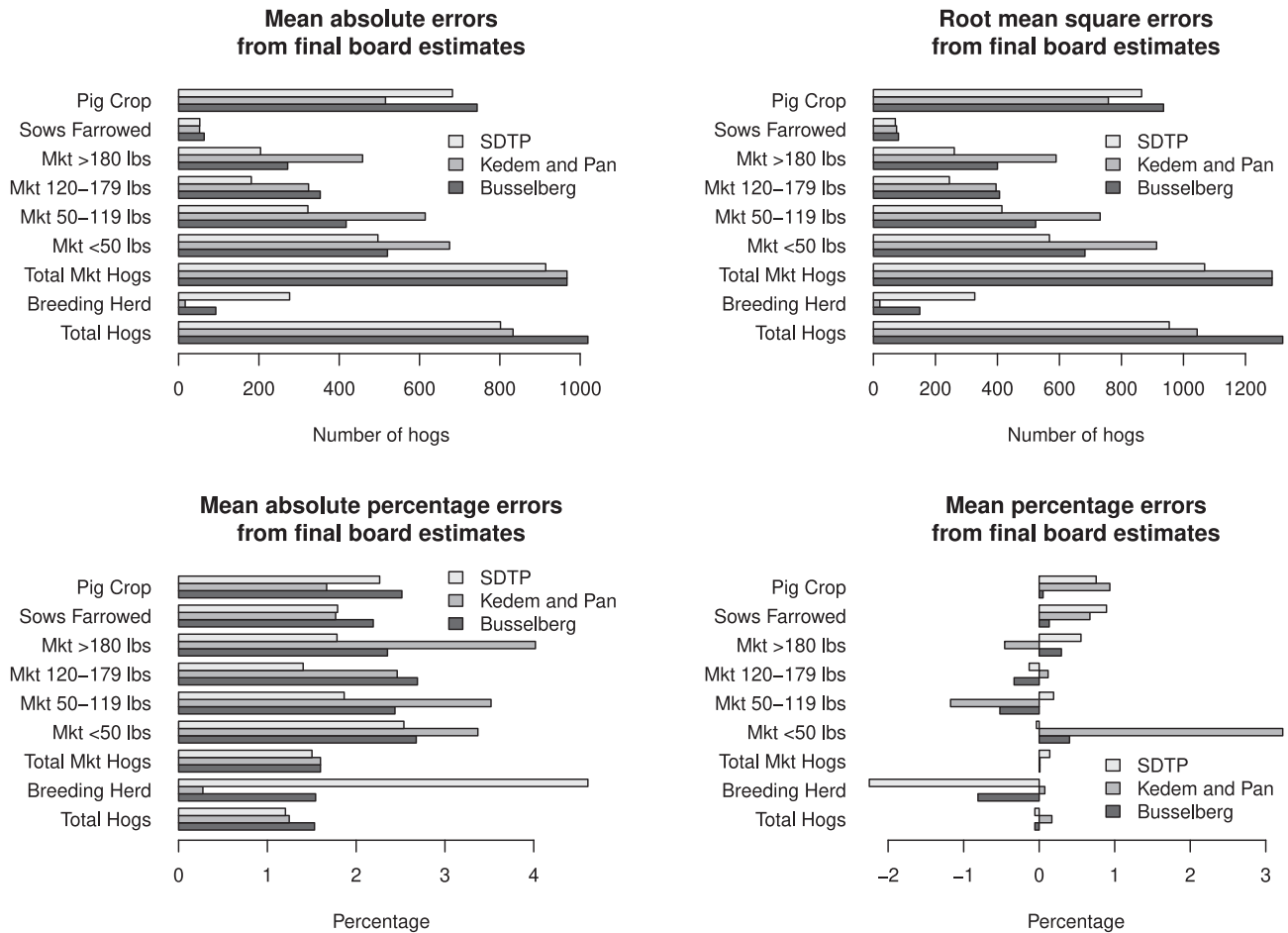
$$\text{MPE}_k = \frac{100\%}{T} \sum_{t=1}^T \left( \frac{y_{k,t}^* - \hat{y}_{k,t}}{y_{k,t}^*} \right). \quad (10)$$

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

NASS official estimates<sup>1</sup> published for the quarters between 2013 and 2017 are used for comparing the results produced by the three models. NASS historical data 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  $\hat{y}_{k,t}$ , for  $k = 3, \dots, 6$ . Total hogs in the US are computed by adding the number of breeding sows and boars to the total value of market hogs.

The comparisons among the SDTP, Kedem and Pan (KP) model, and the SSM are shown in Figure 5. These graphics show the statistics as formulated in Equations (7)–(10). Overall, the KP model produces more accurate results for pig crop, sows farrowed, and breeding herd. The other two models are more reliable in providing results for the four classes of market hogs, and this implies a chain effect on the accuracies of total market hogs and total hogs due to error propagation. The SDTP model produces better inventory numbers for the four market groups, since it accounts for more realistic dynamics of the US swine population. On the other hand, due to the lack of monthly data for market hogs and breeding herd inventory, the SDTP model relies on the SARIMA forecasts of the sows farrowed to quantify the size of the breeding

<sup>1</sup> Fianl official NASS estimates from 2013 to 2017 are available at <https://downloads.usda.library.cornell.edu/usda-esmis/files/jd472w45t/h128nn160/m613n493n/hgpgsb19.pdf> (accessed on August 12, 2020).



**FIGURE 5** Accuracy of the model-based estimates when compared to the final board estimates

herd. This results in a MAPE that is still within 5%, but it is much higher than the MAPEs computed for other variables. Both the KP and the SDTP models perform better than the SSM for non-inventory items. Although the KP model produces the best MAE and RMSE for pig crop, the SDTP model has a similar performance, and both are substantially better than the SSM.

When computing the total number of hogs, the KP model performs better than the SSM due to the high accuracy associated with the breeding herd results. However, the SSM and the KP model have similar MAEs and RMSEs for total market hogs. On the other hand, the MAPEs computed with the results from the SDTP model for the classes of market hogs weighing above 50 lbs are all below 2%. The SDTP model produced a MAPE of 2.54% for market hogs weighing less than 50 lbs. Both the SSM and the KP model produce MAPEs for the four weight classes that are larger than 2.35%. Moreover, the KP model tends to severely overestimate the number of market hogs below 50 lbs and underestimate the market hogs between 50 and 119 lbs.

## 8 | FINAL REMARKS AND CONCLUSIONS

The proposed model produces monthly estimates for the US swine population. All methods currently in use at USDA NASS produce quarterly inventory estimates. This attempt to downscale the time resolution of the swine inventory estimates to the monthly level has been partly successful in producing accurate estimates. In comparison with the current SSM,<sup>1</sup> the SDTP model addresses the rigidity issues due to an over-constrained formulation of the SSM. The SDTP model also benefits from a smoothed formulation of the survival rates to capture departures from periods of equilibrium.

This methodology can be extended to produce state-level estimates that account for interstate transport. The quarterly survey does not provide this information, but other administrative 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. However, it is not clear how to incorporate web-scraped information in the modeling process. The current state of this technology provides warnings related to disease outbreaks affecting the hog population.

## ACKNOWLEDGEMENTS

This manuscript reports statistical methodology developed for the U.S. Department of Agriculture, National Agricultural Statistics Service (NASS), as a collaboration with the National Institute of Statistical Sciences (NISS). The findings and conclusions in this publication are those of the authors and should not be construed to represent any official USDA, or US Government determination or policy.

## DATA AVAILABILITY STATEMENT

The record-level data used in this study have been collected by the National Agricultural Statistics Service. The disclosure of such data is in violation of US laws.

## ORCID

Luca Sartore  <https://orcid.org/0000-0002-0446-1328>

## REFERENCES

1. Busselberg S. The use of signal filtering for hog inventory estimation. Paper presented at: Proceedings of the Federal Committee on Statistical Methodology (FCSM) Research Conference; 2013.
2. Busselberg S. Bridging livestock survey results to published estimates through state-space models: a time series approach. Proceedings of the Government Statistics Section, Joint Statistical Proceedings; 2011.
3. Kedem B, Pan L. *Time Series Prediction of Hog Inventory*. Washington, DC: Unpublished Internal Document, United States Department of Agriculture NASS; 2015.
4. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J Royal Stat Soc Ser B (Methodol)*. 1977;39(1):1-22.
5. Bartlett M. *Stochastic Population Models in Ecology and Epidemiology*. Methuen's Monographs on Applied Probability and Statistics. London: Methuen and Co.; 1960.
6. Shull C. *Modeling Growth of Pigs Reared to Heavy Weights* [PhD thesis]. University of Illinois at Urbana-Champaign, Champaign, IL; 2013.
7. Park H, Spann K, Whitley N, Oh S. Comparison of growth performance of Berkshire purebreds and crossbreds sired by Hereford and Tamworth breeds raised in alternative production system. *Asian Australas J Anim Sci*. 2017;30(9):1358-1362. <https://doi.org/10.5713/ajas.16.0987>.
8. Park H, Oh S. Seasonal variation in growth of Berkshire pigs in alternative production systems. *Asian Australas J Anim Sci*. 2017;30(5):749-754. <https://doi.org/10.5713/ajas.16.0587>.
9. Key N, McBride WD. The changing economics of US hog production. USDA-ERS Economic Research Report 52; 2007.
10. Bridges T, Turner L, Smith E, Stahly T, Loewer O. A mathematical procedure for estimating animal growth and body composition. *Trans ASAE*. 1986;29(5):1342-1347.
11. Liz E, Pilarczyk P. Global dynamics in a stage-structured discrete-time population model with harvesting. *J Theor Biol*. 2012;297:148-165.
12. Box G, Jenkins G, Reinsel G, Ljung G. *Time Series Analysis: Forecasting and Control*. Hoboken, NJ: John Wiley & Sons; 2015.
13. Tibshirani R. Regression shrinkage and selection via the lasso. *J Royal Stat Soc Ser B (Methodol)*. 1996;58(1):267-288.
14. Wang H, Li G, Tsai CL. Regression coefficient and autoregressive order shrinkage and selection via the lasso. *J Royal Stat Soc Ser B (Stat Methodol)*. 2007;69(1):63-78.
15. Box G, Jenkins G. *Time Series Analysis: Forecasting and Control*. San Francisco, CA: Holden-Day Press; 1970.
16. Pollard J. On the use of the direct matrix product in analysing certain stochastic population models. *Biometrika*. 1966;53(3-4):397-415.
17. Eilers P, Marx B. Flexible smoothing with B-splines and penalties. *Stat Sci*. 1996;11(2):89-102.
18. NASS. *Estimation Manual Volume 4: Livestock and Dairy*. Washington, DC: Unpublished Internal Document, United States Department of Agriculture NASS; 2005.
19. Fletcher R. *Practical Methods of Optimization*. Hoboken, NJ: Wiley-Interscience Publication; Wiley; 1987.
20. Byrd R, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained optimization. *SIAM J Sci Comput*. 1995;16(5):1190-1208. <https://doi.org/10.1137/0916069>.
21. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. *Stat Surv*. 2010;4:40-79.
22. Guntuboyina A, Lieu D, Chatterjee S, Sen B. Adaptive risk bounds in univariate total variation denoising and trend filtering. *Ann Stat*. 2020;48(1):205-229.
23. Ver Hoef J. Who invented the delta method? *Am Stat*. 2012;66(2):124-127. <https://doi.org/10.1080/00031305.2012.687494>.
24. Fan J, Wang W, Zhong Y. Robust covariance estimation for approximate factor models. *J Econ*. 2019;208(1):5-22.



25. Cape J, Tang M, Priebe CE. Signal-plus-noise matrix models: eigenvector deviations and fluctuations. *Biometrika*. 2019;106(1):243-250.
26. Hyndman R, Koehler A. Another look at measures of forecast accuracy. *Int J Forecast*. 2006;22(4):679-688.
27. Khan A, Hildreth W. *Case Studies in Public Budgeting and Financial Management*. New York, NY: Marcel Dekker; 2003.
28. Givens G, Hoeting J. *Computational Statistics (Wiley Series in Computation Statistics)*. Hoboken, NJ: Wiley; 2005.

**How to cite this article:** Sartore L, Wei Y, Abayomi E, et al. Modeling swine population dynamics at a finer temporal resolution. *Appl Stochastic Models Bus Ind*. 2020;36:1060–1079. <https://doi.org/10.1002/asmb.2597>

## APPENDIX A. SIMULATION STUDY

This appendix provides additional information on the SDTP model developed at USDA NASS to quantify the size of the US swine population on a monthly basis. A simulation study based on a parametric bootstrap technique is conducted with survey and historical data to assess both the uncertainties of the parameters and the stability of the SDTP model (by looking at the fitted values). The bootstrap replicates are produced in two steps:

1. The survey data are adjusted and aggregated at the national level to form the so-called ADXX estimates. These values are further adjusted and calibrated to remove systematic bias.
2. To produce pre-board estimates, the SDTP model uses the adjusted survey estimates as response values for the current and the previous two months. Historical data are also used to inform the model on earlier temporal dynamics (see Figure 4).

In contrast to the other models developed at NASS, the SDTP model uses the information available from the quarterly survey on a monthly basis. The inventory numbers, such as breeding herd and market hogs (distinguished in four weight groups) refer to quantities measured on the first day of the month when the survey is conducted. Data points on a monthly basis are generated by rearranging the information through the allocation of monthly pig crop and sows farrowed on that corresponding month. The inventory numbers appear on data points at the first day of the month when the survey is conducted, but they are missing for the two months before data collection (see Figure 4). The missing values are then imputed using the model dynamics and the information from the pig crop and sows farrowed monthly data, which are fully available.

For this simulation study, December 2017 is considered as the current month. Parametric bootstrap<sup>28</sup> is used to generate random numbers drawn from a normal distribution with mean and variance computed directly from the ADXX survey data of the December 2017. The random draws replace the original aggregated survey data obtained for the last quarter of 2017 (i.e., for time  $t$ ,  $t - 1$  and  $t - 2$ ; see Figure 4), while the historical values for time  $t - j$ , where  $j \in \{3, 4, \dots\}$ , remain fixed throughout the simulation. The monthly results produced by a single bootstrap replicate are then processed to resemble the quarterly statistics presented for the pre-board (see Table A1). Repeated estimates are produced using the SDTP model by processing 111 quarterly data and 120 rearranged data points on a monthly basis. For each bootstrap replicate, the fitted values of the inventories, pig crop and sows farrowed are stored together with the estimates of the model parameters allowing the properties of the proposed model and estimation methodology to be studied.

The averages, standard errors, and coefficients of variation (CV) shown in Table A1 for inventory and non-inventory items are produced with 1000 bootstrap replicates. These model results indicate that the regression methodology developed for the SDTP is very stable, in fact the standard errors and CV are small.

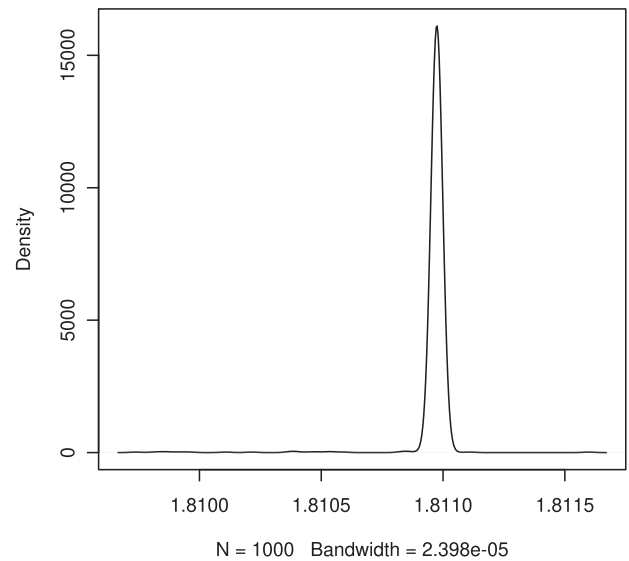
The SDTP parameters can be grouped according to their purpose into four classes:

1. Farrowing rate ( $\varphi$ ). The bootstrap mean is computed to be around 1.811 with 0.0001 standard deviation. Figure A1 shows the empirical distribution of the bootstrap estimates.
2. SARIMA dynamics for (pig crop and sows farrowed). Mean and standard error of parameter estimates are shown in Table A2, while the bootstrap estimate distribution are represented by the box-plots in Figure A2.
3. Survival rates ( $\zeta_t$ ). The average dynamics and their standard errors computed with the 1000 bootstrap replicates are shown in Figure A3. The time series of the bootstrap means of the estimated survival rates reasonably show a cyclical pattern with a seasonal minimum in the autumn. As expected, the standard errors computed at the extremes of the

**TABLE A1** SDTP model outcomes from 1000 bootstrap replicates

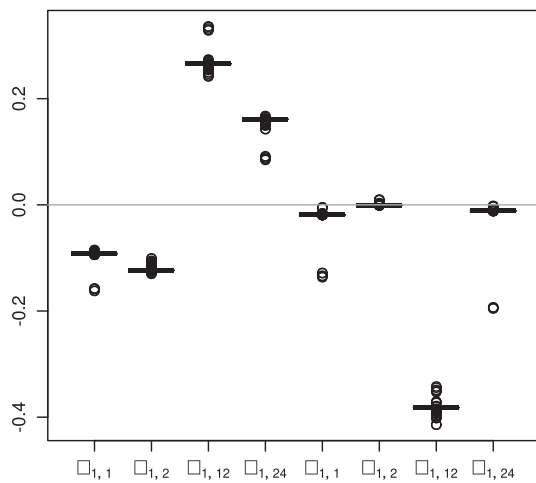
	Average (1000 heads)	Std. Err.	C.V. (%)
Tot. Hogs	69,099.857	13.400	0.019
Breeding Herd	6562.179	1.021	0.016
Market Hogs	62,537.678	13.251	0.021
Group 1	18,744.718	12.063	0.064
Group 2	18,406.682	6.924	0.038
Group 3	12,942.250	5.977	0.046
Group 4	12,444.028	2.794	0.022
Sows Farrowed	3161.076	1.588	0.050
Pig Crop	33,744.931	14.700	0.044
Litter Rate	10.681	0.007	0.066

**Empirical density of farrowing rate estimates**

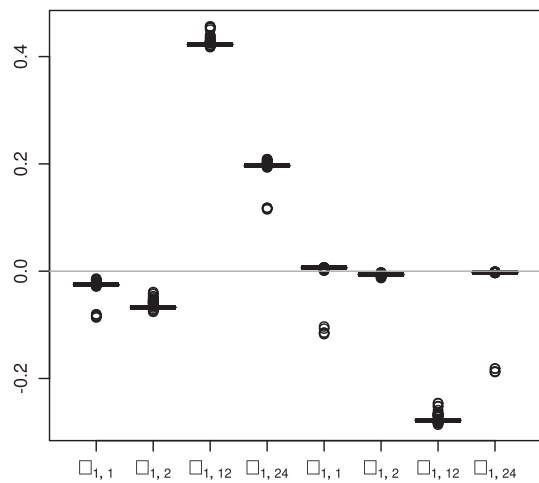


**FIGURE A1** Empirical density function for the farrowing rate estimated with 1000 bootstrap replicates

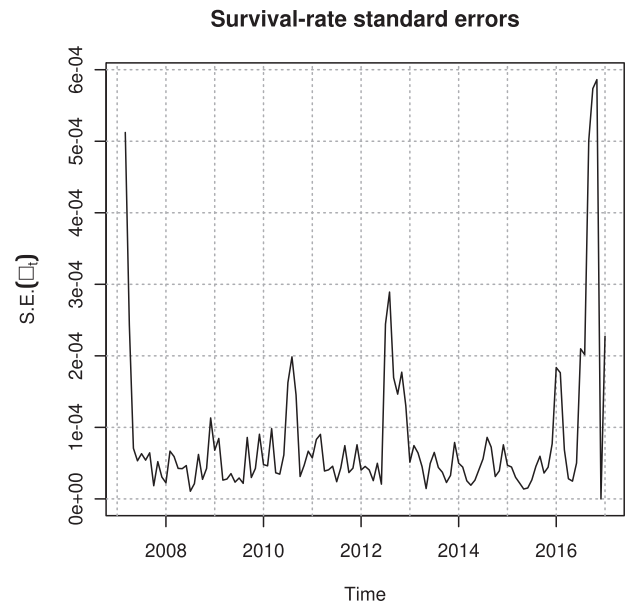
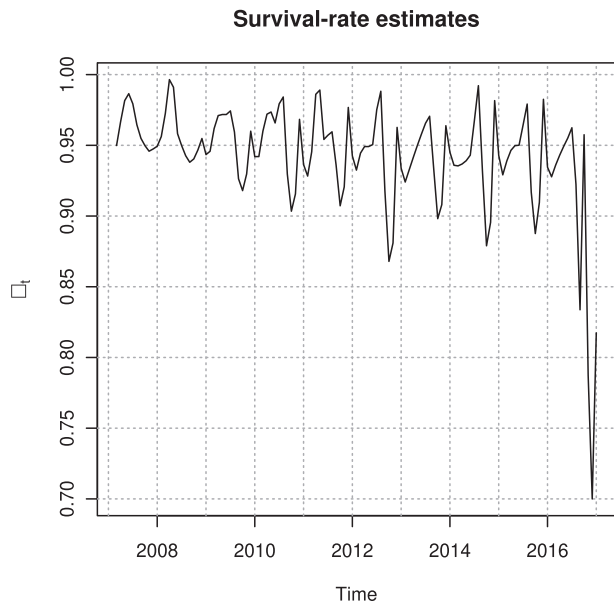
**Parameter estimates for pig-crop model**



**Parameter estimates for sows-farrowed model**

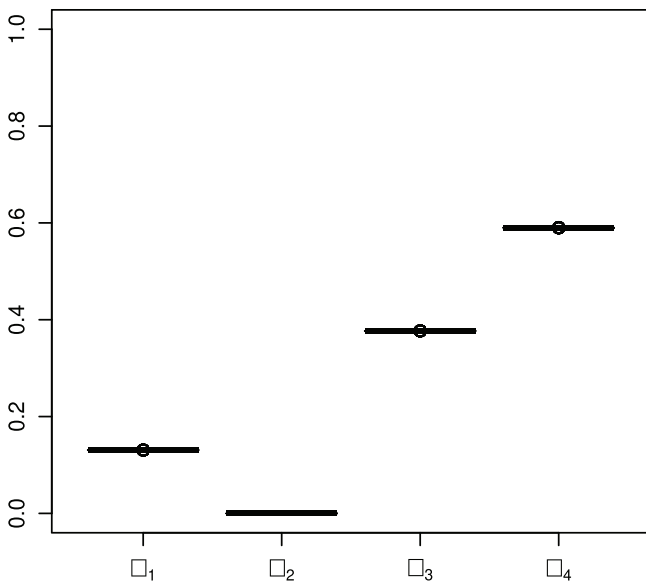


**FIGURE A2** Box-plots for the bootstrap parameter estimates of the SARIMA models for pig crop (on the left) and sows farrowed (on the right)



**FIGURE A3** Average dynamics of the estimated survival rates (on the left), and standard errors of estimated survival rates (on the right)

**Allocation parameter estimates**



**FIGURE A4** Box-plots of cohort allocation parameter bootstrap estimates from a weight class to the next

**TABLE A2** SARIMA parameter estimates from 1000 bootstrap replicates (standard errors in parenthesis)

	$\phi_{k,1}$	$\phi_{k,2}$	$\phi_{k,12}$	$\phi_{k,24}$	$\theta_{k,1}$	$\theta_{k,2}$	$\theta_{k,12}$	$\theta_{k,24}$
Pig crop, $k = 1$	-0.0922 (0.0048)	-0.1234 (0.0018)	0.2660 (0.0050)	0.1605 (0.0053)	-0.0195 (0.0080)	-0.0007 (0.0006)	-0.3820 (0.00330)	-0.0123 (0.0129)
Sows farrowed, $k = 2$	-0.0258 (0.0041)	-0.0671 (0.0022)	0.4229 (0.0025)	0.1962 (0.0057)	0.0057 (0.0083)	-0.0064 (0.0004)	-0.2776 (0.0023)	-0.0032 (0.0129)

time series are higher than the standard errors for the survival rates computed between 2009 and 2016 with most of the values reported below 0.0001. Furthermore, the bootstrap standard errors evaluated at the end of 2012 show increased variability due to the PEDv outbreak.

- Cohort allocation parameters ( $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ ). Figure A4 shows the box-plot of the bootstrap estimates for the cohort allocation parameters, while Table A3 presents the mean and standard error of the bootstrap estimates.

**TABLE A3** Cohort allocation parameter estimates from 1000 bootstrap replicates (standard errors in parenthesis)

$\alpha_1$	$\alpha_2$	$\alpha_3$	$\alpha_4$
0.13	0.00	0.38	0.59
(2.8e-05)	(0.0e+00)	(2.9e-05)	(5.7e-05)

Copyright of Applied Stochastic Models in Business & Industry is the property of John Wiley & Sons, Inc. and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.