

# China's Missing Pigs: Correcting China's Hog Inventory Data Using a Machine Learning Approach

Yongtong Shao, Minghao Li, Dermot Hayes, Wendong Zhang, Tao Xiong, Wei Xie

**Working Paper 20-WP 607**  
August 2020

**Center for Agricultural and Rural Development**  
**Iowa State University**  
**Ames, Iowa 50011-1070**  
**[www.card.iastate.edu](http://www.card.iastate.edu)**

*Yongtong Shao is Professor, Tianjin University of Commerce, Tianjin, China 300134. E-mail: [shaoyt@tjcu.edu.cn](mailto:shaoyt@tjcu.edu.cn).*

*Minghao Li is Assistant Professor, New Mexico State University, Las Cruces, NM 88003. E-mail: [minghao@nmsu.edu](mailto:minghao@nmsu.edu).*

*Dermot Hayes is Charles F. Curtiss Distinguished Professor in Agriculture and Life Sciences, Iowa State University, Ames, IA 50010. E-mail: [dhayes@iastate.edu](mailto:dhayes@iastate.edu).*

*Wendong Zhang is Assistant Professor, Iowa State University, Ames, IA 50010. E-mail: [wdzhang@iastate.edu](mailto:wdzhang@iastate.edu).*

*Tao Xiong is Professor, Huazhong Agricultural University, Wuhan, China 430070. E-mail: [taoxiong@mail.hzau.edu.cn](mailto:taoxiong@mail.hzau.edu.cn)*

*Wei Xie is Graduate Student, Tianjin University of Commerce, Tianjin, China 300134. E-mail: [1962519158@tjcu.edu.cn](mailto:1962519158@tjcu.edu.cn).*

This publication is available online on the CARD website: [www.card.iastate.edu](http://www.card.iastate.edu). Permission is granted to reproduce this information with appropriate attribution to the author and the Center for Agricultural and Rural Development, Iowa State University, Ames, Iowa 50011-1070.

For questions or comments about the contents of this paper, please contact Tao Xiong, [taoxiong@mail.hzau.edu.cn](mailto:taoxiong@mail.hzau.edu.cn) or Wendong Zhang, [wdzhang@iastate.edu](mailto:wdzhang@iastate.edu). For questions about the online replication codes, contact Tao Xiong.

Iowa State University does not discriminate on the basis of race, color, age, ethnicity, religion, national origin, pregnancy, sexual orientation, gender identity, genetic information, sex, marital status, disability, or status as a U.S. veteran. Inquiries regarding non-discrimination policies may be directed to Office of Equal Opportunity, 3410 Beardshear Hall, 515 Morrill Road, Ames, Iowa 50011, Tel. (515) 294-7612, Hotline: (515) 294-1222, email [eooffice@iastate.edu](mailto:eooffice@iastate.edu).

## **China's Missing Pigs: Correcting China's Hog Inventory Data**

### **Using a Machine Learning Approach**

Yongtong Shao, Minghao Li, Dermot Hayes, Wendong Zhang, Tao Xiong, and Wei Xie

*Forthcoming* at **American Journal of Agricultural Economics**

Wendong Zhang and Tao Xiong are the corresponding authors. For questions about the online replication codes, please contact Xiong.

Yongtong Shao is a professor in the Department of Finance at the Tianjin University of Commerce, 409 Guangrong Rd, Beichen District, Tianjin, China 300134, [shaoyt@tjcu.edu.cn](mailto:shaoyt@tjcu.edu.cn)

Minghao Li is an assistant professor in the Department of Economics, Applied Statistics & International Business at New Mexico State University, Domenici Hall 212, Las Cruces, NM, 88003, 575-646-4124, [minghao@nmsu.edu](mailto:minghao@nmsu.edu)

Dermot Hayes is Charles F. Curtiss Distinguished Professor in Agriculture and Life Sciences in the Department of Economics, the Department of Finance, and the Center for Agricultural and Rural Development at Iowa State University, 518 Farmhouse Lane, 568C Heady Hall, Ames, IA 50011, 515-294-6185, [dhayes@iastate.edu](mailto:dhayes@iastate.edu)

Wendong Zhang is an assistant professor in the Department of Economics and the Center for Agricultural and Rural Development at Iowa State University, 518 Farmhouse Lane, 478C Heady Hall, Ames, IA 50011, 515-294-2536, [wdzhang@iastate.edu](mailto:wdzhang@iastate.edu)

Tao Xiong is a professor and Chair in the Department of Agricultural Economics & Management at Huazhong Agricultural University, 1 Shizhishan Rd, Nanhu District, Wuhan, China 430070, [taoxiong@mail.hzau.edu.cn](mailto:taoxiong@mail.hzau.edu.cn)

Wei Xie is a graduate student in the Department of Finance at the Tianjin University of Commerce, 409 Guangrong Rd, Beichen District, Tianjin, China 300134, [1962519158@qq.com](mailto:1962519158@qq.com)

**Acknowledgments:**

Li, Zhang and Hayes gratefully acknowledge support from the USDA National Institute of Food and Agriculture Hatch Project 101,030 and grant 2019-67023-29414, while Xiong acknowledges the support from the National Natural Science Foundation of China (Project No. 71771101). The authors thank the ISU Center for China-US Agricultural Economics and Policy, where Li was a postdoctoral research associate and Shao and Xiong were visiting scholars. The authors also appreciate editing assistance from Nathan Cook, Becky Olson, and Barbara Nordin, and comments from Guiping Hu, Chad Hart, and Kelvin Leibold. Any remaining errors are the authors' responsibility.

# **China's Missing Pigs: Correcting China's Hog Inventory Data**

## **Using a Machine Learning Approach**

**Running Head (50 characters):**

**Correct China's Hog Inventory Data with Machine Learning**

### **Abstract**

Small sample size often limits forecasting tasks such as the prediction of production, yield, and consumption of agricultural products. Machine learning offers an appealing alternative to traditional forecasting methods. In particular, Support Vector Regression has superior forecasting performance in small sample applications. In this article, we introduce Support Vector Regression via an application to China's hog market. Since 2014, China's hog inventory data has experienced an abnormal decline that contradicts price and consumption trends. We use Support Vector Regression to predict the true inventory based on the price-inventory relationship before 2014. We show that, in this application with a small sample size, Support Vector Regression out-performs neural networks, random forest, and linear regression. Predicted hog inventory decreased by 3.9% from November 2013 to September 2017, instead of the 25.4% decrease in the reported data.

**Keywords:** China, machine learning, prediction, pork, support vector regression

**JEL Codes:** Q02, Q13, Q17

Due to data availability, structural change and the biological cycles of agricultural production, forecasting tasks in agricultural economics often involves time series data with limited sample size. The advance of machine learning (ML), broadly defined as computer algorithms that automatically improve performance, offers appealing alternatives to traditional forecasting tools (Storm, Baylis, and Heckelei 2019). Support Vector Regression (SVR) is especially promising for small sample time-series forecasting common in agricultural economics.

This application of SVR is motivated by abnormal trends in China's hog inventory that obscures the understanding of world's largest pork market. In 2014, China's hog inventory began to deviate from a previously stable relationship with prices—inventory numbers went into rapid decline even though prices were high and consumption was stable. We believe this paradox is due to a recent downward bias in China's inventory data; and, we argue that we can quantify this bias by projecting inventory during the problematic period using a previously stable inventory-price relationship. This forecasting task involves a short time series and potentially a large number of predictors, making it a suitable application for SVR. The objective of this article is to: (a) expose a previously unknown downward bias in China's hog and sow inventory and determine when it started; (b) use SVR to quantify the bias and project actual inventory data; and, (c) compare SVR's forecasting performance to neural networks, random forest, as well as ordinary least square regression (OLS).

In recent years, economists have recognized that ML methods have potentially superior forecasting performance. For example, Mullainathan and Spiess (2017) compare several ML methods with OLS regression and find that the former predicts housing prices

more accurately. Bajari et al. (2015) find that several popular ML methods, including the discrete version of SVR, all predict grocery demand more accurately than OLS and logit regressions.

Among ML methods, SVR holds a unique advantage in data analytics with small sample size (Al-Anazi and Gates 2012; Tange et al. 2017) because it optimally determines model complexity by taking sample size into account (Vapnik 2013). Specifically, only a small subset of observations (support vectors) directly contribute to the final prediction, while the entire set of observations influence results indirectly by determining which observations become support vectors. Furthermore, the use of kernel functions in SVR reduces the number of coefficients in non-linear models, making high-dimensional models feasible for small samples. Despite being one of the most popular ML methods (Wu et al. 2008), we are not aware of any application of SVR in journals in agricultural economics.

China is the world's largest pork producer and consumer, and trends in hog production in China have significant implications for the global pork and feed market. The recent decline in China's hog inventory statistics, if true, is enormous. According to data from China's Ministry of Agriculture and Rural Affairs (MOA), hog and sow inventories decreased by 25.4% and 28.9%, respectively, from November 2013 to September 2017. Despite this decrease in inventory, hog prices showed patterns consistent with a typical hog cycle, with price increases and decreases of magnitudes similar to previous cycles. Furthermore, from 2013 to 2017, consumption of domestic pork, as measured using household surveys and adjusted for net imports, was stable.

Testing for structural breaks and forecasting both require identifying potential determinants of hog inventory. The economics of hog production and pork and hog markets are characterized by complex nonlinear dynamics that result from physical production cycles (Chavas and Holt 1991; Holt and Craig 2006) and the way in which pork producers form price expectations and make decisions based on history and projections (Hayes and Schmitz 1987). The predictors used in this study include the past, current, and future prices of piglets, hogs, pork, corn, soybean meal, and commercial feeds.

While there are theoretically valid reasons to include these predictors, whether their inclusion can improve prediction is an empirical question. We use a filtering method to conduct feature selection, an important procedure in ML for selecting independent variables; and, we demonstrate that feature selection substantially improves prediction accuracy. Furthermore, SVR with feature selection out-performs the best specifications of neural networks, random forest, and OLS.

Our projected hog and sow inventories from November 2013 to September 2017 show decreases of 3.9% and 1.1%, respectively, which is much lower than the respective 25.4% and 28.9% decreases in official MOA data. These predictions are bounded by narrow confidence intervals and are robust to using alternative specifications.

The remaining sections provide graphical analyses of the problem, graphical analyses and econometric test results related to the timing of the structural break, the proposed empirical methodology, the projection results, a discussion of three potential reasons for the downward bias, and conclusions.

## **A Graphical Examination of China's Hog Market Data**

In this section, we first discuss why recent trends in hog and sow inventories are likely due to faulty data. We then briefly introduce a long-understood over-reporting problem and explain how to combine the over-reporting and recent bias corrections to obtain a final estimate of China's hog and sow inventories.

### *Abnormal Trends in Recent Inventory Data*

The MOA began publishing monthly inventory data in January 2009 when volatility in the pork market heightened the need for more accurate and frequent hog production statistics.

The MOA reports hog and sow inventories, as measured in percentage change relative to the previous month, based on a sample of hog production facilities selected from 400 counties.

Figure 1 shows MOA hog and sow inventory data and corresponding monthly average hog prices from China Animal Agriculture Association (CAAA). To put the magnitude of the inventory reductions into perspective, from November 2013 to September 2017, MOA's official hog (sow) inventory declined by 25.4% (28.9%), which is equivalent to 161.4% (233.1%) of the total 2017 hog (sow) inventory in the United States.

There are two major data inconsistencies in the MOA hog and sow inventory data. First, the large inventory declines did not have a discernable impact on hog prices, which increased from 11.1 yuan/kg to 20.6 yuan/kg from April 2014 to May 2016, and then decreased after that. The price increase and decrease in this hog cycle are of similar magnitude and duration to those in previous hog cycles (see figure 1).

The inconsistency between inventory and price data in recent years is transparent in the cumulative sum plot (cusum plot) shown in figure 2 below. This shows the sum of residuals from a time-series regression over time (Brown, Durbin, and Evans 1975). When parameters are stable, the expected value of the sum of residuals is zero; and, if parameters change over time, the cumulative sum of residuals will drift away from zero. Figure 2 is the cusum plot for the OLS regression with hog inventory as the dependent variable and hog price as the independent variable. The cumulative sum of residuals is close to zero at the beginning; however, at some point during 2013 and 2014, the sum of residuals starts to drift downwards, eventually becoming statistically different from zero in July 2015.

Second, the relationship between hog inventory and pork consumption is inconsistent. Using pork consumption data reported by the National Bureau of Statistics of China (NBSC 2019) and pork import data from Global Trade Atlas (GTA 2019) we estimate that consumption of domestic pork decreased by only 0.3% from 2013 to 2017. Given that pork storage capacity is negligible compared to production, consumption of pork should approximately equal production; therefore, stable consumption is at odds with the large decline in reported hog inventory.

### *The Over-reporting Problem*

Researchers have long recognized that China's hog production data are over-reported, as local officials inflate the economic performance of their jurisdictions (Lu 1998; Lohmar 2015). Yu and Abler (2014), Ma, Huang, and Rozelle (2004), and Fuller, Hayes, and Smith

(2000) propose various methods to deflate China's hog production; and, studies on China's pork production routinely acknowledge or correct for over-reporting bias (e.g., Wang et al. 2013; Jin et al. 2010; Rae et al. 2006).

The MOA only reports monthly changes in inventory while users of the data series multiply monthly changes to the initial level of inventory (2008) published by NBSC to obtain levels. Therefore, the over-reporting problem in NBSC inventory in 2008 will inflate the entire series by some constant.

Yu and Abler (2014) represent the most recent and comprehensive correction for China's over-reporting of pork production data. They document that since 1996, NBSC has been the only authorized agency to publish national statistics, which explains why MOA only publishes monthly changes instead of levels. Yu and Abler's (2014) adjustment strategy recognizes that China's hog industry consists of traditional backyard production and an emerging commercial hog sector. They estimate backyard-sector per capita pork production using household survey data and multiply that number by rural population to get total backyard production; and, they estimate commercial pork production by dividing commercial feed used by a feed conversion ratio. They then measure total production by adding backyard production to commercial production. Yu and Abler's (2014) best estimate shows that the true level of pork production in 2008 is 78% of reported production, which is the baseline year for MOA's data series. Although their correction is for pork production, we assume the same over-reporting rate for hog inventories, since NBSC's weight of 75.73 kg per reported pig is realistic (Yu and Abler 2014, table 2). We also assume the same over-reporting rate for sow inventory, since the ratio between hog inventory and sow inventory (about 10:1) is also

realistic in our data. Therefore, we multiply the entire hog and sow data series by a factor of 0.78. We correct the level after adjustment of the aberrant inventory data, but we can reverse the sequence with no effect on the final results.

### **Data and Testing the Structural Break in the Inventory-Price Relationship**

We identify the structural break in the inventory-price relationship using the supremum Wald test for an unknown single structural break (Andrews 1993; also see Perron 2006 for a survey of related methods). At each possible date, hog and sow inventories are repeatedly regressed on prices (table 1) with different sets of parameters before and after the break date. The date with the best goodness-of-fit is determined to be the date of the structural break. As robustness checks, we experiment with alternative models with different lag and lead terms (table 2). The supremum Wald test is used to evaluate whether the highest Wald statistic is statistically different from what is expected from a data series with no structural break. We perform a search for the break date between July 2013 and July 2015 based on our graphical analysis in figures 1 and 2. Table 1 presents the summary statistics for all data we use in this study.

Table 2 presents specifications and results for the supremum Wald test. These models vary by including only contemporaneous prices ( $t$ ), three recent lagged prices ( $t-1$  to  $t-3$ ), three long lags ( $t-10$  to  $t-12$ ), or three lead prices ( $t+1$  to  $t+3$ ). All price data are from CAAA (2019). Table 2 shows that all supremum Wald tests reject the null hypothesis of no structural break in our study period; and, the test narrows the range of the break point to between October 2013 and February 2015 (see table 2). To avoid using problematic inventory data in

the training/testing dataset for the SVR prediction, we choose the earliest structural break point (October 2013) as the last data point in the training/testing period.<sup>1</sup> Table 2 presents the summary statistics of the training and testing data and the data used for our SVR projections.

## **Methodology**

This section reviews the literature on ML and SVR, introduces the SVR method, explains the general procedure for feature selection, model training and testing, then presents the setup for comparison models including neural networks, random forest, and OLS.

### *Machine Learning and SVR in the Literature*

Arthur Samuel (1959) first coined the term machine learning, which refers to computer algorithms that improve automatically through experience. Economists use ML for prediction in estimating productivity (Chalfin et al. 2016), policy evaluation (McBride and Nichols 2018), and testing theory (Peysakhovich and Naecker 2017). Recently, economists have begun using ML for causal identification (Athey 2019). Agricultural economists use neural networks to predict farmers' risk preferences (Kastens and Featherstone 1996), count the number of federal regulations (Malone and Chambers 2017), and forecast commodity prices (Dharmasena, Bessler, and Capps Jr. 2016; Ribeiro and Oliveira 2011). Er (2018) uses several ML algorithms to predict irrigated farmland prices in Kansas. For reviews of ML in economics, see Mullainathan and Spiess (2017), Ghoddusi, Creamer, and Rafizadeh (2019), Athey (2019), and Athey and Imbens (2019); and, for reviews of ML in agricultural economics see Woodard (2016) and Storm, Baylis, and Heckelei (2019).

The discrete version of SVR, called support vector machine (SVM), originated from the statistical learning theory developed by Vapnik and Chervonenkis (1974; see Vapnik 2013 for a textbook treatment in English), which spells out the discrepancy between training errors and prediction errors. This discrepancy tends to increase with the complexity of the model and decrease with larger sample size. Therefore, if model complexity is not regulated, small sample prediction is prone to overfitting (i.e. small training error and large testing/prediction error). SVM, and its close relative SVR, aim to minimize prediction error and provide an edge in higher dimension and smaller sample prediction. A variety of applications in engineering and medical research empirically demonstrate the superiority of SVM and SVR in small sample prediction.<sup>2</sup> However, previous research has also shown SVM or SVR may not be suitable for large data sets due to computational tractability (Ho and Lin 2012).

Economists use SVR in the field of energy economics to predict electricity demand (Hahn, Meyer-Nieberg, and Pick 2009; Li et al. 2012) and electricity prices (Mirakyan, Meyer-Renschhausen, and Koch 2017). Financial researchers have used SVR to predict corporate bond recovery rates (Nazemi, Heidenreich, and Fabozzi 2018) and stock prices (Henrique, Sobreiro, and Kimura 2018). Agricultural researchers have used SVM in crop yield estimation and livestock, water, and soil management (Liakos et al. 2018) and carcass weight prediction for beef cattle (Alonso, Castañón, and Bahamonde 2013). Liu et al. (2019) use SVR in the prediction of hog prices, Jheng, Li, and Lee (2018) use it to predict rice yield, and Huang (2015) uses it to evaluate agricultural project bids; however, the application of SVR is absent from leading agricultural economics journals.

## *Introduction to SVR*

In general, we can write the relationship between the dependent variable ( $y$ ) and vector of predictors ( $\mathbf{x}$ ) as:

$$(1) \quad y = \boldsymbol{\beta}\boldsymbol{\varphi}(\mathbf{x}) + \beta_0$$

where  $\boldsymbol{\varphi}(\mathbf{x})$  is some non-linear transformation of  $\mathbf{x}$  (e.g., a polynomial or translog function);  $\boldsymbol{\beta}$  is a vector of parameters; and,  $\beta_0$  is the intercept. We first discuss the linear case in equation (2), then extend the model to the non-linear case in equation (1).

$$(2) \quad y = \boldsymbol{\beta}\mathbf{x} + \beta_0$$

The first objective of SVR is to fit data points into a belt formed by two lines (or hyperplanes in the multivariate case) with fixed  $y$ -direction distance ( $\varepsilon$ ) from the regression line, making the regression line as flat as possible (see figure 3). Intuitively, all else equal, a flatter regression line means less influence from noise in the predictor ( $\mathbf{x}$ ). We achieve this objective by minimizing the norm of the slope vector ( $\|\boldsymbol{\beta}\|$ ). For mathematical convenience, we represent this objective as  $\frac{1}{2}\|\boldsymbol{\beta}\|^2$  in the minimization problem: if  $\frac{1}{2}\|\boldsymbol{\beta}\|^2$  is minimized, then  $\|\boldsymbol{\beta}\|$  is minimized.

Another intuitive interpretation for the above objective is that, with fixed  $\varepsilon$ , a flatter regression line means a wider  $\varepsilon$ -belt (the gap perpendicular to the regression line defines the belt width), which better captures out-of-sample data points. This is based on a basic geometrical relationship (figure 3): the width  $d = 2\varepsilon \cdot \cos\theta$ , where  $\cos\theta$  is larger when the regression line is flatter. This is apparent from comparing the two panels in figure 3: with the same  $\varepsilon$ , the flatter regression line on the left makes the  $\varepsilon$ -belt wider.

Given a fixed  $\varepsilon$ , it is not always possible to fit all data points within the  $\varepsilon$ -belt. The second objective of SVR is to allow some data points to fall outside of the  $\varepsilon$ -belt while minimizing the sum of their absolute distances ( $|\zeta_i|$ ) to the edge of the  $\varepsilon$ -belt along the  $y$ -axis. This objective is the counterpart of minimizing sum of squares in OLS. In contrast to OLS, only some data points (those outside the  $\varepsilon$ -belt) are penalized: due to the constraint in the minimization problem, the smallest  $|\zeta_i|$  is zero for points within the  $\varepsilon$ -belt, since  $|y_i - \boldsymbol{\beta}\mathbf{x} - \beta_0| \leq \varepsilon$ . The minimization problem for SVR represents the two competing objectives above. The tuning parameter ( $C$ ) governs the relative weight given to the second objective.

$$\min \frac{1}{2} \|\boldsymbol{\beta}\|^2 + C \sum_{i=1}^N |\zeta_i|$$

$$w.r.t. \quad \beta_0, \boldsymbol{\beta}, \zeta_i$$

$$subject\ to \quad |y_i - \boldsymbol{\beta}\mathbf{x} - \beta_0| \leq \varepsilon + |\zeta_i|$$

The minimization problem readily shows the first reason SVR is good for small sample prediction—only observations outside of the  $\varepsilon$ -belt (called support vectors) contribute to the objective function, while data points within the  $\varepsilon$ -belt exert indirect effects by determining the set of support vectors. The second reason lies in the extension from linear SVR to non-linear SVR. The dual of the above minimization problem (see Smola and Schölkopf (2004) for details) shows that optimal coefficients only depend on the inner products of support vectors  $\mathbf{x}_i\mathbf{x}_j$ . The extension from linear-SVR to non-linear SVR is achieved by replacing  $\mathbf{x}_i\mathbf{x}_j$  with  $\boldsymbol{\varphi}(\mathbf{x}_i)\boldsymbol{\varphi}(\mathbf{x}_j)$ , where  $\boldsymbol{\varphi}(\mathbf{x}_i)$  is some non-linear transformations of  $\mathbf{x}_i$  with a potentially large number of coefficients (see equation (1)). Using the “kernel trick” (Boser, Guyon, and Vapnik 2003), we can approximate the inner product  $\boldsymbol{\varphi}(\mathbf{x}_i)\boldsymbol{\varphi}(\mathbf{x}_j)$  with certain kernel functions  $k(\mathbf{x}_i, \mathbf{x}_j)$  with small and fixed number of

parameters. The Radial Basis Function kernel  $k(x_i, x_j, \gamma) = \exp(-\frac{\|x_i - x_j\|}{\gamma})$  is one of the most commonly used kernel functions (Tange et al. 2017). The ability to reduce the number of coefficients using the kernel trick is another reason why SVR performs well with small sample size—it allows non-linearity without costing too many degrees of freedom.

### *Feature Selection, Training, and Testing*

Prediction tasks using ML, regardless of the specific modeling technique, usually consist of training, testing, and prediction. In our case, we split the dataset into a training sample, a testing sample, and a prediction sample. We use the first 46 observations, from January 2009 to October 2012, for training; and, we use the 12 observations, from November 2012 to October 2013, for testing. We optimize feature and parameter selection for different models (SVR, neural network, random forest, and OLS) on the training sample, and evaluate prediction performance on the testing sample. Finally, we perform forecasting from November 2013 to September 2017 using the selected models.

The training step consists of feature selection and parameter selection. For feature selection, economists usually select variables based on the underlying economic theory; however, it is often the case that competing theories suggest different inputs, or that the theoretical effect of a variable is ambiguous. Feature selection can improve prediction accuracy by selecting independent variables with high predicative power (Han, Pei, and Kamber 2011), which is especially important when the number of observations is relatively small compared to the number of variables (Sorjamaa et al. 2007), as in our case. Variables

can either be selected a priori based on certain criteria (the filter method), or selected based on their realized performance (the wrapper method).

Our application starts from a host of predictors (see table 1) that include the monthly prices of soybean meal, corn, composite commercial feed, piglets, live hogs, and pork with various lag/lead variables. We include lagged prices up to three months to capture delayed producer response to prices and/or strategic response based on recent price trends. Lead prices up to three steps ahead capture reverse causation of inventories on prices and producers' strategic anticipation of future prices. Price lags of 10–12 months capture the physical production cycle from sow breeding to hog slaughter. While these variables are plausible inputs, their effects can be ambiguous (Hayes and Schmitz 1987), hence the need for feature selection.

We adopt the filter method with a commonly used criteria based on mutual information (MI) (Sorjamaa et al. 2007). Compared to the simple correlation measure, MI can detect all dependency, whereas correlation can only detect linear dependency. For example, in the bivariate case, if  $y = \sin(x) + e$ , the correlation between  $x$  and  $y$  will be zero, MI will not. We select a 50% subset of predictors that have the highest MI with the dependent variables.<sup>3</sup>

Another aspect of training is to select parameters with the optimal predictive performance. For SVR, several parameters need tuning—the weight on the penalty function ( $C$ ), the belt in which the penalty is zero ( $\epsilon$ ), and parameters in the kernel function. We optimize these parameters based on out-of-sample prediction performance using the ten-fold cross-validation method in the training process. After training, we compare alternative

models, including SVR with RBF kernel, SVR with linear kernel, neural networks, random forest, and OLS, in the testing process. We use three error measures: root mean squared error (RMSE), normalized mean squared error (NMSE), and mean absolute percentage error (MAPE) to evaluate out-of-sample prediction accuracy.<sup>4</sup>

### *Comparison of Models*

We compare the forecasting performance of SVR with random forest, neural network, and OLS. In all models, we have the same price variables with the same lag/lead structures all normalized to 0~1. SVR is implemented using the *LIBSVM* package (version 3.1) in Matlab (Chang and Lin 2011).

The OLS models need to address non-stationarity in the data. Augmented Dickey-Fuller tests (available upon request) show that all independent variables are stationary after first differencing; however, dependent variables are only stationary at the 5% significance level after second differencing. Therefore, all variables are in second difference in the OLS models.

Random forest is a supervised ML algorithm that operates by constructing a multitude of decision trees at the training process and producing the mean prediction of the individual trees. Random forest is increasingly popular because it can cope with higher-order interactions and even highly correlated predictor variables (Strobl et al. 2008). Considering our small sample size, we limit our experiment to six possible trees (a ten-fold cross-validation method determines the best number of trees)—5, 10, 15, 20, 25, 30. We implement the random forest using the *TreeBagger* function in Matlab.

Neural network models offer several advantages, including the ability to detect complex non-linear relationships and the availability of multiple structure and training algorithms (Tu 1996). However, researchers also dub it a black box for its low interpretability of input features, susceptibility to over-fitting, low calculation robustness, and significant training time (Tu 1996). We implement a three-layer feed-forward neural network (FNN) with fully connected nodes in adjacent layers. The three-layer FNN consists of the input layer, the hidden layer, and the output layer. We use hidden nodes with nonlinear activation functions to process the information received by the input nodes. The Levenberg and Marquardt algorithm is used for training. As for the architecture of FNN, price variables determine the number of input nodes and the number of output nodes is set to one, denoting the predicted value of hog or sow inventory. We choose the number of hidden nodes from 15, 20, 25, 30, 35, and 40 using ten-fold cross-validation. We implement FNN using the *feedforwardnet* function in Matlab.

## **Results**

Tables 3 and 4 show the out-of-sample prediction performance of various models for hog and sow inventories, respectively. Comparing results across input choices within each prediction method demonstrates the importance of formal feature selection in ML—no matter which metrics we use, specifications identified by feature selection always perform better than ad hoc input choices. Relative to the average of other input choices in tables 3 and 4, feature selection reduces the prediction error for the hog (sow) inventories by 52%~79 (57%~84%) for SVR, 52%~80% (23%~47%) for the neural networks, and 11%~21% (17%~41%) for the

random forest. Interestingly, feature selection provides no benefit for the OLS method, with most ad hoc variable choices performing better than the specification chosen by feature selection in both hog and sow inventory prediction.

Comparing across methods, the SVR model with the best specification substantially outperforms the other models with their best specifications. For hog (sow) inventory prediction, depending on metrics of prediction accuracy, the prediction error for best specification of SVR is 28%~48% (33%~55%) less than that of neural networks, 58~85% (58%~82%) less than that of random forest, and 9%~21% (12%~23%) less than that of OLS. When the linear kernel is used instead of the RBF kernel in SVR, the prediction error is 119%~427% (39%~114%) greater (results available upon request). The importance of feature selection is again prominent—for ad hoc variable choices, SVR is only the best performing model in a minority of cases, and OLS often performs best. It is likely that a more flexible functional form would further increase the performance of the OLS model. However, given the limited number of observations in the training and testing dataset, there is little room to increase the dimensionality of the OLS model. In fact, there are not enough degrees of freedom to include all of the lead and lag terms in the OLS model. This is why SVR's ability to handle small-sample, high-dimensional problems is useful for applications such as ours.

The underlying assumption for our correction is that the actual relationships between inventory and prices are stable throughout the entire training, testing, and forecasting periods. Based on diagnostic evidence from the cusum plot, we are reasonably confident that the inventory-price relationship is stable before the data break. We are not aware of any event after the structural break that would have caused such a drastic decline in inventory.

Figures 4a and 4b show the projected hog and sow inventories based on the specification suggested by feature selection. Figure 4a shows that from November 2013 to September 2017, the projected hog inventory decreased by 3.9% instead of by 25.4%, as MOA data indicate. The gap between the SVR projection and the actual data at the end of the projection period is 99 million head, or 28.2% of the reported data. Narrow confidence intervals calculated using the bootstrapping method (Lins et al. 2015) bound the projection. The projected sow inventory shown in figure 4b is also substantially higher than the observed data. From November 2013 to September 2017, the results show that the sow inventory decreased by 1.1% rather than by 28.9%, as in official MOA data. By the end of the projection, the difference between projected and observed data is 11 million sows, or 28.2% of the reported data.

The lower lines in figures 4a and 4b show the inventory levels before and after the structural break if we correct both the newly discovered underreporting bias using SVR after the structural break and address the over-reporting bias prior to the structural break, using the correction proposed by Yu and Abler (2014). Results suggest that the two biases now approximately cancel out and that the current MOA inventory data are close to being correct. To evaluate the degree of model uncertainty, we present the range of predictions spanned by the top five SVR specifications, as measured by RMSE for hog and sow inventories, respectively. Figures 5a and 5b show that different specifications produce similar predictions.

### **Possible Reasons for the Bias in Recent Inventory Trends**

The first and preferred explanation is that officials corrected for over-reporting ahead of China's Third National Agricultural Census. The over-reporting rate in 2008 is 28%, which is very close to the difference between our predicted hog inventory and the official statistics at the end of the period. Thus, it is possible that using hand-held tablets that uploaded results to a central system without intermediaries, an innovation in the third census (Chen 2016), prompted local officials to deflate data in advance. This explanation is consistent with the adjustments made after each census. While the first and second census led to downward corrections in pork production statistics by 21.8% and 7.2%, respectively, the third census led to a slight upward adjustment of 2.4%.<sup>5</sup>

A second possible explanation for the reporting bias is that the pressure of increasingly stringent environmental regulations led local producers and government officials to gradually under-report hog inventory. China enacted a new environmental protection law in 2014, which increased penalties for environmental violations (Li and Frederick 2015). In 2016, MOA announced the thirteenth five-year plan for agriculture and made moving hog production away from waterways and urban population centers a major policy goal (Inouye 2017). In various provinces in urban southeast China, MOA forbids hog operations in certain areas. In this new policy environment, both producers and local governments may have the incentive to under-report hog inventory in order to meet environmental goals imposed by upper-level governments. If this is true, we would expect the underreporting problem to be more serious in regions with more stringent environmental regulations. Unfortunately, we do not have provincial inventory data series to assess this possibility.

While environmental regulations may cause underreporting, they are unlikely to reduce actual inventory. First, a large decrease in actual inventory is inconsistent with the pattern of prices and consumption described earlier. Moreover, the purpose of the environmental policy is to transfer hog production away from environmentally sensitive regions, not to reduce overall hog production. In fact, the provinces in which MOA is enforcing environmental controls (the “development control zone”) only account for 35% of the total inventory in 2013. Since MOA is encouraging production in other regions—the “development focus zone,” the “moderate development zone,” and the “potential development zone”—the overall impact of environmental policy on actual inventory is likely small.

A third possible explanation is sampling bias caused by the rapid spatial reconfiguration of China’s hog production. MOA designed the environmental policies to shift pork production away from environmentally sensitive regions. As previously discussed, we base the changes in hog inventory on a sample of 400 major hog production counties; and, it is likely that in recent years, production has shifted away from some of these counties. In March 2018, MOA revised inventory change data for the previous month downward by 0.5%, citing statistical bias caused by the redistribution of hog production. Even if we assume this downward sampling bias occurred every month since the end of 2013, we arrive at hog and sow inventories that are roughly 5% to 9% lower than projected inventories, which suggests that this bias alone may not be enough to explain the observed underreporting. If the statistical sampling is indeed biased, we would expect regions with increasing production to have more under-reporting, because our data do not capture these new facilities.

## Summary and Conclusions

In this article, we introduce SVR, an ML method that is especially suitable for small sample prediction because it can automatically adjust model complexity according to sample size.

We demonstrate SVR's small sample performance by comparing it to random forest, neural networks, and OLS. With proper feature selection, SVR consistently out-performs the other three methods. Small sample predictions are very common in agricultural economics, and SVR can be a valuable addition to an economist's toolbox. While SVR is good for small sample prediction, it is computationally expensive with large datasets and other ML methods may be more appropriate.

Our research sheds light on the true state of China's hog industry. Researchers have long been aware of the over-reporting problem in China's pork production and routinely apply downward corrections in analyses. Recently, a new data problem has emerged in which a substantial decrease in inventory contradicts a normal price cycle and stable consumption of domestic pork. The new problem compounds with the old over-reporting problem and further obscures the true state of China's pork production. Uncertainty about China's hog inventory hinders the assessment of important events, such as the recent African Swine Fever outbreak (Global AgriTrends 2019).

After correcting for the new downward bias and over-reporting in the base year, we estimate hog and sow inventories to be 351.3 million and 38.8 million head, respectively, in September 2017—close to the 349.5 million and 35.4 million head in the official data. This

demonstrates that the downward reporting bias between September 2013 and September 2017 offset the over-reporting at the beginning of the period.

We identify three possible reasons for this downward bias in the inventory data—under-reporting to deflate data before China’s Third National Agricultural Census, under-reporting due to pressure from stringent environmental regulations, and sampling bias caused by rapid geographical shifts in hog production. We believe that the first of these explanations is the most likely.

## References

- Al-Anazi, A.F. and I.D. Gates. 2012. "Support Vector Regression to Predict Porosity and Permeability: Effect of Sample Size." *Computers & Geosciences* 39:64–76.
- Alonso, J., Á.R. Castañón, and A. Bahamonde. 2013. "Support Vector Regression to Predict Carcass Weight in Beef Cattle in Advance of the Slaughter." *Computers and Electronics in Agriculture* 91:116–20.
- Andrews, D.W.K. 1993. "Tests for Parameter Instability and Structural Change with Unknown Change Point." *Econometrica: Journal of the Econometric Society* 64(4):821–56.
- Athey, S. 2019. "The Impact of Machine Learning on Economics: An Agenda." In A.K. Agrawal, J. Gans, and A. Goldfarb, eds. *The Economics of Artificial Intelligence: An Agenda*. Chicago, IL: University of Chicago Press, pp. 507–47.
- Athey, S. and G.W. Imbens. 2019. "Machine Learning Methods That Economists Should Know About." *Annual Review of Economics* 11(1):685–725.
- Bajari, P., D. Nekipelov, S.P. Ryan, and M. Yang. 2015. "Machine Learning Methods for Demand Estimation." *American Economic Review: Papers and Proceedings* 105(5):481–85.
- Boser, B.E., I.M. Guyon, and V.N. Vapnik. 2003. "A Training Algorithm for Optimal Margin Classifiers." In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, 144–52.
- Brown, R.L., J. Durbin, and J.M. Evans. 1975. "Techniques for Testing the Constancy of Regression Relationships Over Time." *Journal of the Royal Statistical Society: Series B*

3(2):149–63.

Chalfin, A., O. Danieli, A. Hillis, Z. Jelveh, M. Luca, J. Ludwig, and S. Mullainathan. 2016.

"Productivity and Selection of Human Capital with Machine Learning." *American Economic Review* 106(5):124–27.

Chang, C.-C., and C.-J. Lin. 2011. "LIBSVM: A Library for Support Vector Machines."

*ACM Transactions on Intelligent Systems and Technology* 2(3):1–27. (Version 3.1 most recently updated in November 2019. Software available at <https://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html>).

Chavas, J.-P., and M.T. Holt. 1991. "On Nonlinear Dynamics: The Case of the Pork Cycle."

*American Journal of Agricultural Economics* 73(3):819–28.

Chen, W. 2016. "Ten Years Later, the Third Agricultural Census Comes." Xinhua.Net. 2016.

[http://www.xinhuanet.com//politics/2016-12/15/c\\_1120125583.htm](http://www.xinhuanet.com//politics/2016-12/15/c_1120125583.htm).

China Animal Agriculture Association. 2019. "Animal Product Price Reports."

<http://www.caaa.cn/>.

Dharmasena, S., D.A. Bessler, and O. Capps Jr. 2016. "Food Environment in the United

States as a Complex Economic System." *Food Policy* 61:163–75.

Doerr, B., P. Fischer, A. Hilbert, and C. Witt. 2017. "Detecting Structural Breaks in Time

Series via Genetic Algorithms." *Soft Computing* 21(16):4707–20.

Er, E. 2018. "Applications of Machine Learning to Agricultural Land Values: Prediction and

Causal Inference." PhD dissertation, Kansas State University.

Fuller, F., D. Hayes, and D. Smith. 2000. "Reconciling Chinese Meat Production and

Consumption Data." *Economic Development and Cultural Change* 49(1):23–43.

- Ghoddusi, H., G.G. Creamer, and N. Rafizadeh. 2019. "Machine Learning in Energy Economics and Finance: A Review." *Energy Economics* 81:709–27.
- Global AgriTrends. 2019. "The China ASF Puzzle." *International Meat Market Update* 13(5):1.
- Global Trade Atlas (GTA). 2019. <https://www.gtis.com/gta/>.
- Golland, P., W.E.L. Grimson, M.E. Shenton, and R. Kikinis. 2000. "Small Sample Size Learning for Shape Analysis of Anatomical Structures." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 72–82.
- Hahn, H., S. Meyer-Nieberg, and S. Pickl. 2009. "Electric Load Forecasting Methods: Tools for Decision Making." *European Journal of Operational Research* 199(3):902–7.
- Han, J., J. Pei, and M. Kamber. 2011. *Data Mining: Concepts and Techniques*. New York: Elsevier.
- Hayes, D.J., and A. Schmitz. 1987. "Hog Cycles and Countercyclical Production Response." *American Journal of Agricultural Economics* 69(4):762–70.
- Henrique, B.M., V.A. Sobreiro, and H. Kimura. 2018. "Stock Price Prediction Using Support Vector Regression on Daily and up to the Minute Prices." *The Journal of Finance and Data Science* 4(3):183–201.
- Ho, C.-H. and C.-J. Lin. 2012. "Large-Scale Linear Support Vector Regression." *Journal of Machine Learning Research* 13(Nov):3323–48.
- Holt, M.T. and L.A. Craig. 2006. "Nonlinear Dynamics and Structural Change in the U.S. Hog—Corn Cycle: A Time-Varying STAR Approach." *American Journal of Agricultural Economics* 88(1):215–33.

- Huang, M. 2015. "Agricultural Economic Evaluation Based on Improved Support Vector Regression." In *2015 8th International Conference on Intelligent Computation Technology and Automation (ICICTA)*, 118–21.
- Inouye, A. 2017. *Chinese Consumers Substitute Burgers For Bacon In 2017*. Washington, DC: US Department of Agriculture, GAIN report no. CH17005.
- Jheng, T.-Z., T.-H. Li, and C.-P. Lee. 2018. "Using Hybrid Support Vector Regression to Predict Agricultural Output." In *2018 27th Wireless and Optical Communication Conference (WOCC)*, 1–3.
- Jin, S., H. Ma, J. Huang, R. Hu, and S. Rozelle. 2010. "Productivity, Efficiency and Technical Change: Measuring the Performance of China's Transforming Agriculture." *Journal of Productivity Analysis* 33(3):191–207.
- Kastens, T.L. and A.M. Featherstone. 1996. "Feedforward Backpropagation Neural Networks in Prediction of Farmer Risk Preferences." *American Journal of Agricultural Economics* 78(2):400–415.
- Li, D., C. Chang, C. Chen, and W. Chen. 2012. "Forecasting Short-Term Electricity Consumption Using the Adaptive Grey-Based Approach—An Asian Case." *Omega* 40(6):767–73.
- Li, W., and C. Frederick. 2015. *China's Increasing Appetite for Imported Beef*. Washington, DC: U.S. Department of Agriculture, GAIN report no. CH15034.
- Liakos, K.G., P. Busato, D. Moshou, S. Pearson, and D. Bochtis. 2018. "Machine Learning in Agriculture: A Review." *Sensors* 18(8):2674.
- Lins, I.D., E.L. Droguett, M. das Chagas Moura, E. Zio, and C.M. Jacinto. 2015. "Computing

- Confidence and Prediction Intervals of Industrial Equipment Degradation by Bootstrapped Support Vector Regression." *Reliability Engineering & System Safety* 137:120–28.
- Liu, C., and Y. Cheng. 2018. "An Application of the Support Vector Machine for Attribute-By-Attribute Classification in Cognitive Diagnosis." *Applied Psychological Measurement* 42(1):58–72.
- Liu, Y., Q. Duan, D. Wang, Z. Zhang, and C. Liu. 2019. "Prediction for Hog Prices Based on Similar Sub-Series Search and Support Vector Regression." *Computers and Electronics in Agriculture* 157:581–88.
- Lohmar, B. 2015. "Will China Import More Corn?" *Choices* 30(2):1–7.
- Lu, F. 1998. "What Are the Real Production and Consumption Data for Meat, Egg and Aquatic Products in China?" In *CCER Discussion Paper C1998005*. Peking University Beijing, China.
- Ma, H., J. Huang, and S. Rozelle. 2004. "Reassessing China's Livestock Statistics: An Analysis of Discrepancies and the Creation of New Data Series." *Economic Development and Cultural Change* 52(2):445–73.
- Malone, T., and D. Chambers. 2017. "Quantifying Federal Regulatory Burdens in the Beer Value Chain." *Agribusiness* 33(3):466–71.
- McBride, L., and A. Nichols. 2018. "Retooling Poverty Targeting Using Out-of-Sample Validation and Machine Learning." *The World Bank Economic Review* 32(3):531–50.
- Ministry of Agriculture and Rural Affairs of the People's Republic of China (MOA). 2019. Hog Inventory Data From 400 Monitoring Counties. <http://www.moa.gov.cn/>.

- Mirakyan, A., M. Meyer-Renschhausen, and A. Koch. 2017. "Composite Forecasting Approach, Application for Next-Day Electricity Price Forecasting." *Energy Economics* 66:228–37.
- Mullainathan, S., and J. Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31(2):87–106.
- National Bureau of Statistics of China (NBSC). 2019. China Statistical Yearbooks. 2019. <http://www.stats.gov.cn/english/Statisticaldata/AnnualData/>.
- Nazemi, A., K. Heidenreich, and F.J. Fabozzi. 2018. "Improving Corporate Bond Recovery Rate Prediction Using Multi-Factor Support Vector Regressions." *European Journal of Operational Research* 271(2):664–75.
- Perron, P. 2006. Dealing With Structural Breaks. *Palgrave Handbook of Econometrics* 1(2):278–352.
- Peysakhovich, A. and J. Naecker. 2017. "Using Methods From Machine Learning to Evaluate Behavioral Models of Choice Under Risk And Ambiguity." *Journal of Economic Behavior & Organization* 133:373–84.
- Rae, A.N., H. Ma, J. Huang, and S. Rozelle. 2006. "Livestock in China: Commodity-Specific Total Factor Productivity Decomposition Using New Panel Data." *American Journal of Agricultural Economics* 88(3):680–95.
- Ribeiro, C.O., and S.M. Oliveira. 2011. "A Hybrid Commodity Price-Forecasting Model Applied to the Sugar-Alcohol Sector." *Australian Journal of Agricultural and Resource Economics* 55(2):180–98.
- Samuel, A.L. 1959. "Some Studies in Machine Learning Using the Game of Checkers." *IBM*

- Journal of Research and Development* 3(3):210–29.
- Smola, A.J., and B. Schölkopf. 2004. "A Tutorial on Support Vector Regression." *Statistics and Computing* 14(3):199–222.
- Sorjamaa, A., J. Hao, N. Reyhani, Y. Ji, and A. Lendasse. 2007. "Methodology for Long-Term Prediction of Time Series." *Neurocomputing* 70(16–18):2861–69.
- Storm, H., K. Baylis, and T. Heckelei. 2019. "Machine Learning in Agricultural and Applied Economics." *European Review of Agricultural Economics*. URL: <https://doi.org/10.1093/erae/jbz033>.
- Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. 2008. "Conditional Variable Importance for Random Forests." *BMC Bioinformatics* 9(1):307.
- Tange, R.I., M.A. Rasmussen, E. Taira, and R. Bro. 2017. "Benchmarking Support Vector Regression against Partial Least Squares Regression and Artificial Neural Network: Effect of Sample Size on Model Performance." *Journal of Near Infrared Spectroscopy* 25(6):381–90.
- Tu, J.V. 1996. "Advantages and Disadvantages of Using Artificial Neural Networks Versus Logistic Regression for Predicting Medical Outcomes." *Journal of Clinical Epidemiology* 49(11):1225–31.
- Vapnik, V. 2013. *The Nature of Statistical Learning Theory*. New York, NY: Springer Science & Business Media.
- Vapnik, V.N., and A.Y. Chervonenkis. 1974. "The Method of Ordered Risk Minimization, I. (in Russian)" *Avtomatika i Telemekhanika* 8:21–30.
- Wang, S.L., F. Tuan, F. Gale, A. Somwaru, and J. Hansen. 2013. "China's Regional

- Agricultural Productivity Growth in 1985–2007: A Multilateral Comparison." *Agricultural Economics* 44(2):241–51.
- Woodard, J.D. 2016. "Data Science And Management For Large Scale Empirical Applications in Agricultural and Applied Economics Research." *Applied Economic Perspectives and Policy* 38(3):373–88.
- Wu, X., V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J. Hand, and D. Steinberg. 2008. "Top 10 Algorithms in Data Mining." *Knowledge and Information Systems* 14(1):1–37.
- Xing, F., and P. Guo. 2004. "Classification of Stellar Spectral Data Using SVM." In *International Symposium on Neural Networks*, 616–21.
- Yu, X., and D. Abler. 2014. "Where Have All the Pigs Gone? Inconsistencies in Pork Statistics in China." *China Economic Review* 30:469–84.
- Yu, Y., T. McKelvey, and S.Y. Kung. 2013. "A Classification Scheme for ‘High-Dimensional-Small-Sample-Size’ Data Using Soda And Ridge-SVM With Microwave Measurement Applications." In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 3542–46.

## Grouped Endnotes

<sup>1</sup> The evolutionary ML algorithm, proposed by Doerr et al. (2017), is an alternative to the supremum Wald test. The evolutionary ML method identifies December 2014 as the structural break date for hog inventory, and November 2014 for sow inventory.

<sup>2</sup> For examples in engineering see Al-Anazi and Gates (2012), Tange et al. (2017), Xing and Guo (2004), and Yu, McKelvey, and Kung (2013). For applications in medical research see Golland et al. (2000), and Liu and Cheng (2018).

<sup>3</sup> The equation for bivariate MI is:

$$MI(X, Y) = \int_{-\infty}^{+\infty} \rho^{X,Y}(x, y) \log \frac{\rho^{X,Y}(x, y)}{\rho^X(x)\rho^Y(y)} dx dy$$

where  $\rho^X(x)$  and  $\rho^Y(y)$  are the pdf of random variables  $x$  and  $y$ , and  $\rho^{X,Y}(x, y)$  is the joint pdf of  $x$  and  $y$ . Estimating MI involves the estimation of these density functions.

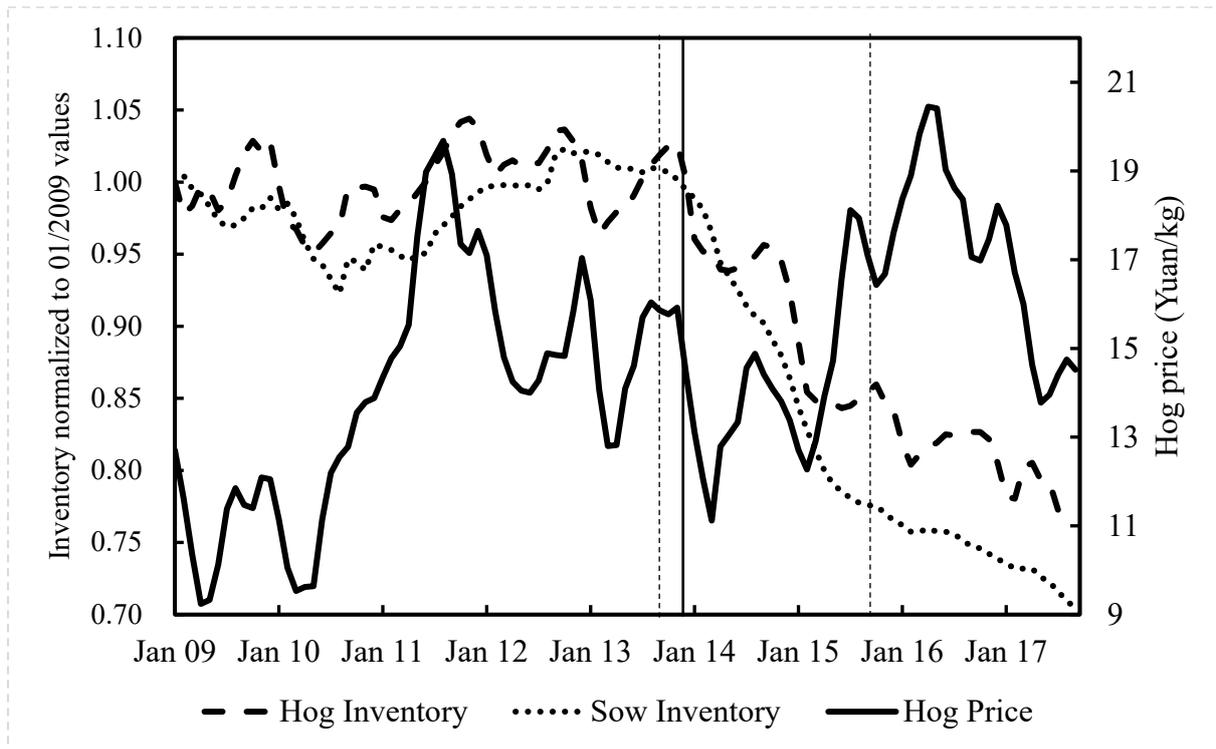
$$^4 \text{ RMSE} = \sqrt{\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N}},$$

$$\text{NMSE} = \frac{1}{N} \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{\left(\frac{1}{N} \sum_{i=1}^N y_i\right) \left(\frac{1}{N} \sum_{i=1}^N \hat{y}_i\right)},$$

$$\text{MAPE} = \frac{100}{N} \sum_{i=1}^N \frac{|y_i - \hat{y}_i|}{y_i}, \text{ where } N \text{ is the number of observations, } y_i \text{ is the observed}$$

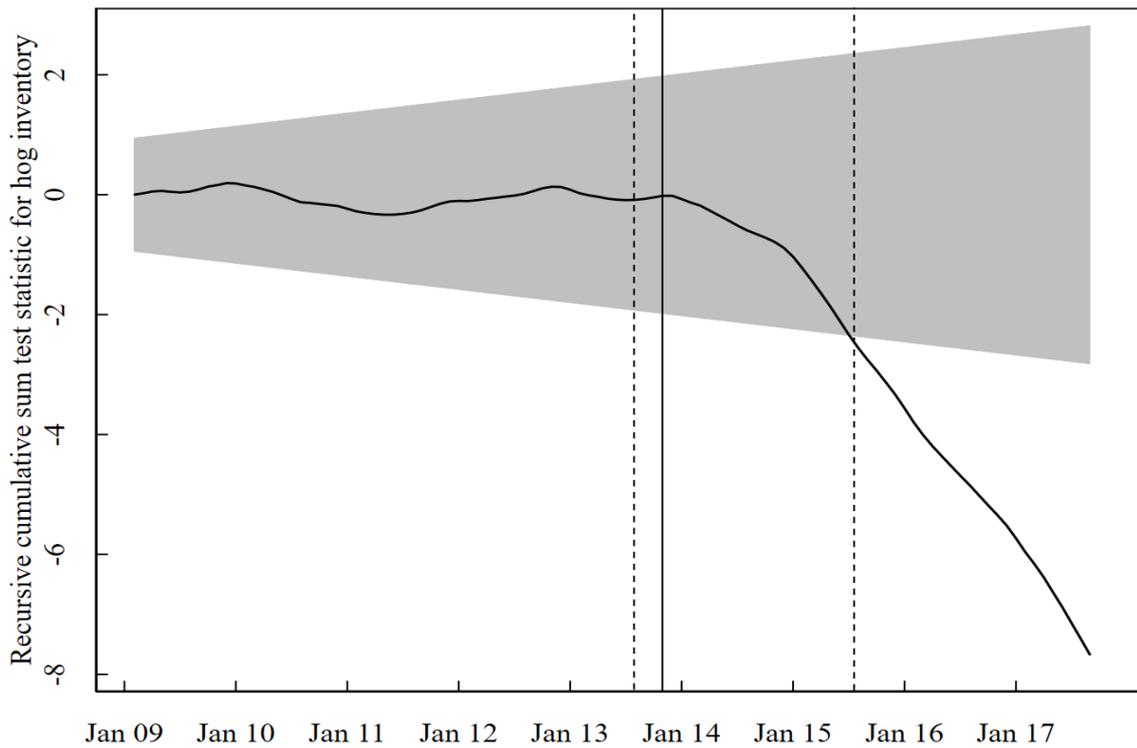
value, and  $\hat{y}_i$  is the predicted value.

<sup>5</sup> Adjustments are authors' calculations based on changes in data for the same year across various issues of China Statistical Yearbooks.



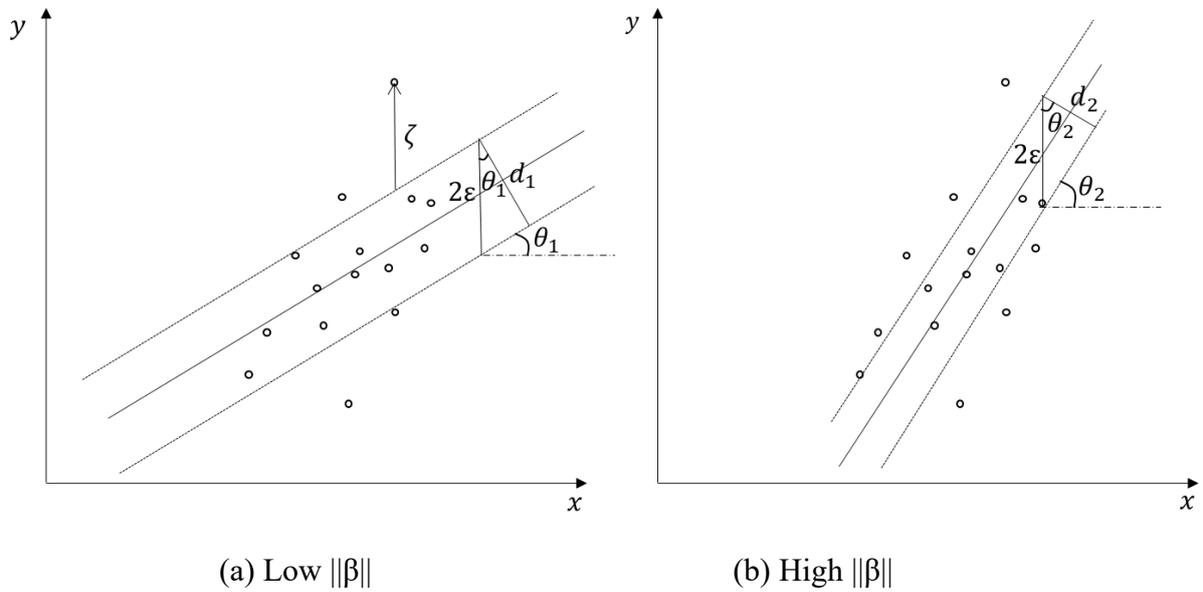
**Figure 1. China’s hog and sow inventories and hog price, January 2009 to September 2017**

*Note:* The dashed vertical lines indicate the range (July 2013 to July 2015) we use to search for a structural break and the solid vertical line represents the structural break date used to separate the training and testing period from the projection period.



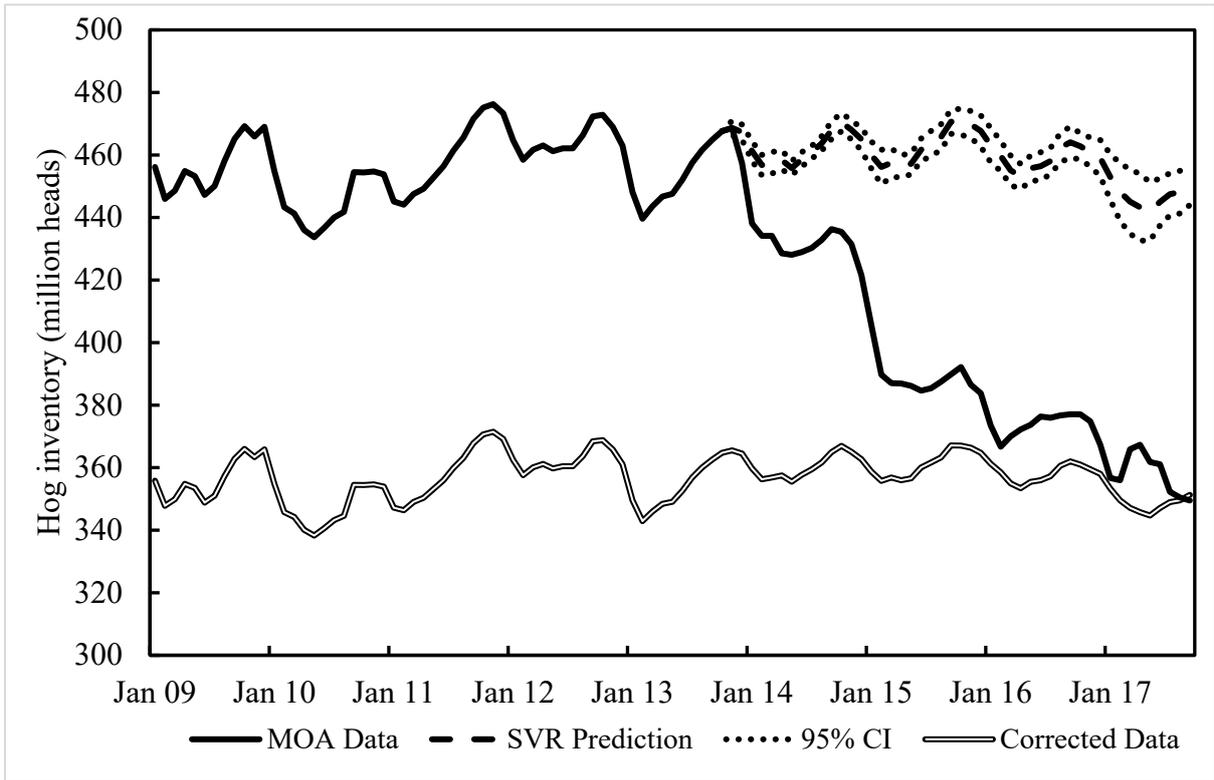
**Figure 2. Cumulative sum of recursive errors**

*Note:* We calculate recursive errors from an OLS regression of hog inventory on hog prices (Brown, Durbin, and Evans 1975). The gray area represents the 95% confidence interval for the null hypothesis of the cumulative sum of recursive errors equaling zero. The dashed vertical lines indicate the range (July 2013 to July 2015) we use to search for a structural break and the solid vertical line represents the structural break date used to separate the training and testing period from the projection period.

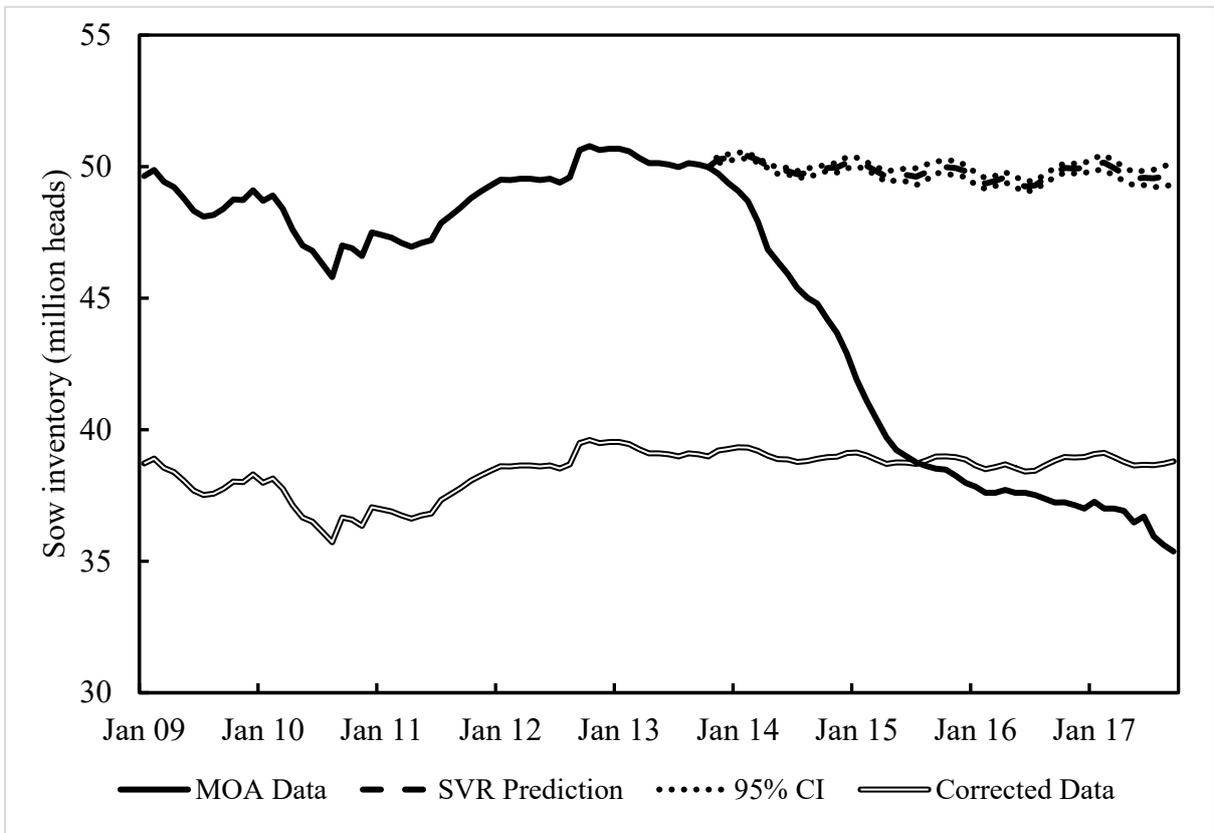


**Figure 3. SVR as visualized in a one-dimensional linear case**

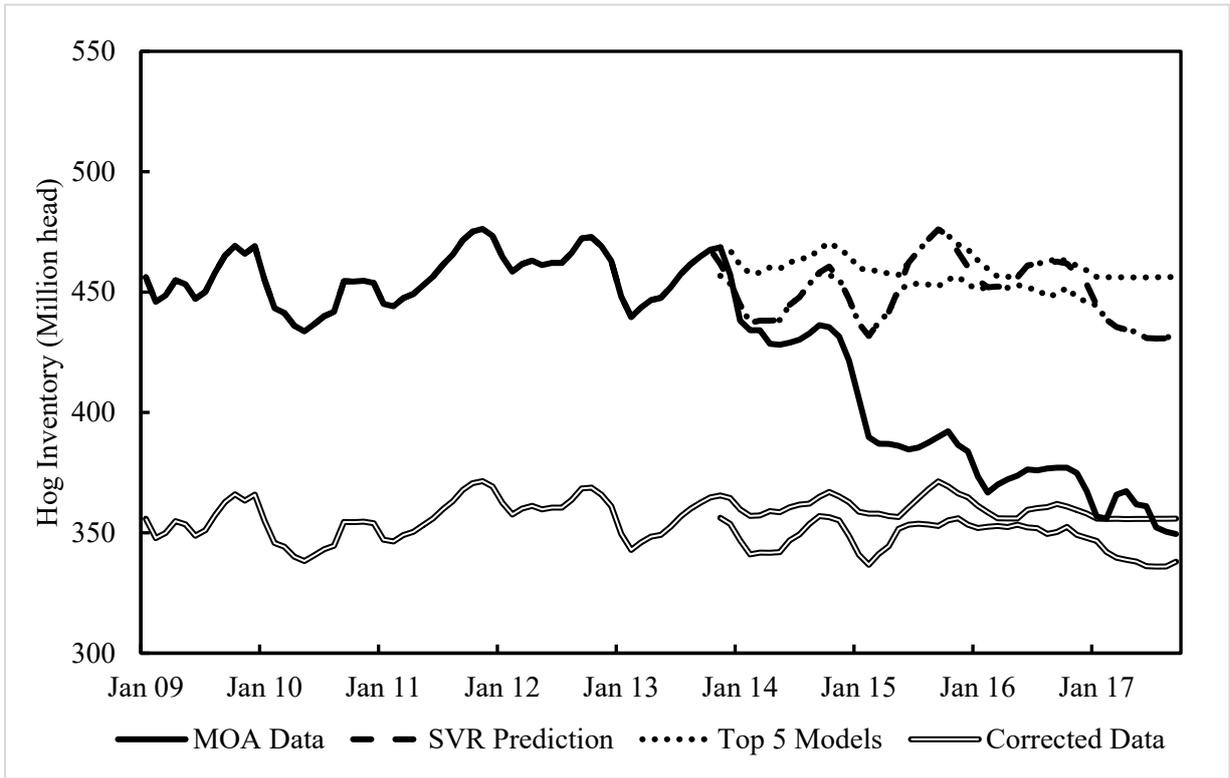
*Note:* In these figures,  $\varepsilon$  is the y-direction distance from the edge of the  $\varepsilon$ -belt to the regression line and  $d_1$  and  $d_2$  are the perpendicular distances between the two edges of the  $\varepsilon$ -belt. Data points outside the  $\varepsilon$ -belt are punished according to vertical distance ( $\zeta$ ) to the edge of the  $\varepsilon$ -belt. Both panels have the same  $\varepsilon$ , but the  $\varepsilon$ -belt in the left panel is wider ( $d_1 > d_2$ ) because the slope is flatter. If the fitness ( $\sum_{i=1}^N |\zeta_i|$ ) of the regression lines is the same in both panels, SVR will favor the left panel.



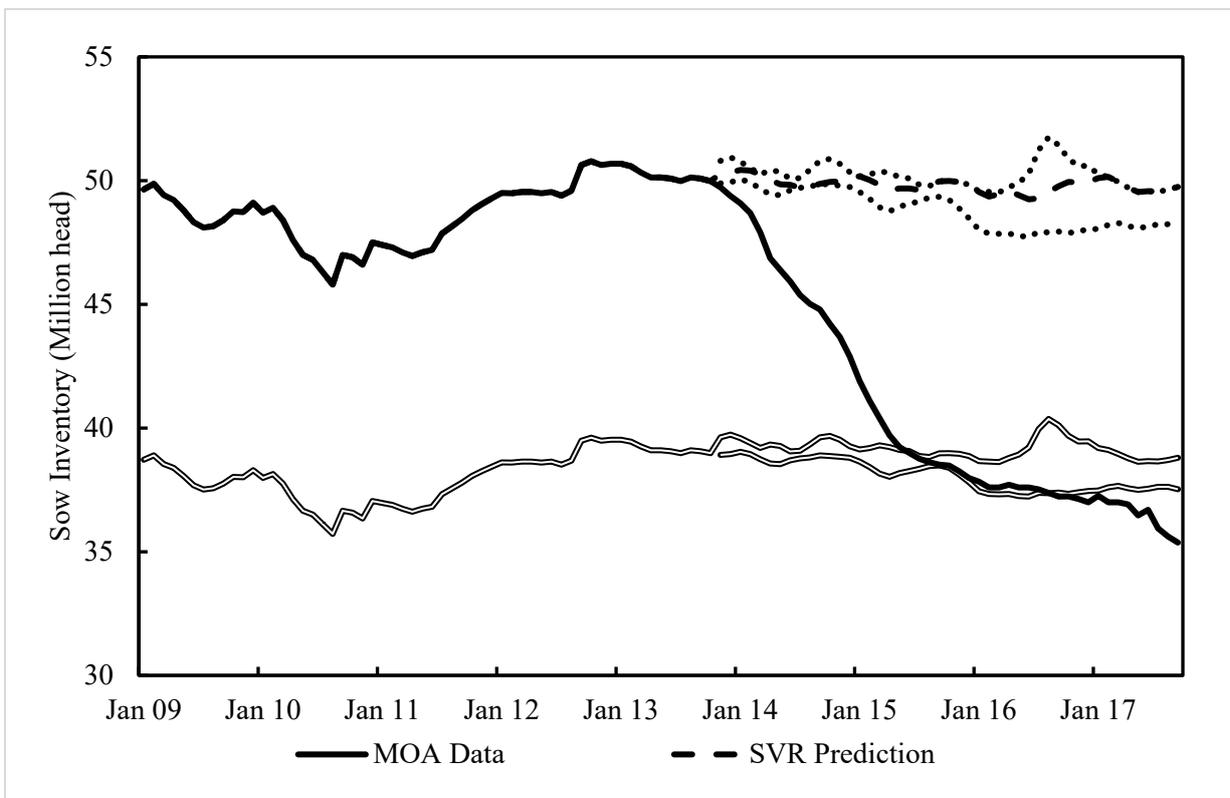
**Figure 4a. Hog inventory prediction**



**Figure 4b. Sow inventory prediction**



**Figure 5a. Model uncertainty for the hog inventory projection**



**Figure 5b. Model uncertainty for the sow inventory projection**

**Table 1. Summary Statistics**

| Variable name                    | Training & testing data |                    | Projection data   |                    |
|----------------------------------|-------------------------|--------------------|-------------------|--------------------|
|                                  | (01/2009–10/2013)       |                    | (11/2013–09/2017) |                    |
|                                  | Mean                    | Standard deviation | Mean              | Standard deviation |
| MOA hog inventory (million head) | 456.1                   | 10.9               | 393.0             | 31.2               |
| MOA sow inventory (million head) | 48.8                    | 1.3                | 40.4              | 4.3                |
| Soybean meal price (yuan/kg)     | 3.8                     | 0.3                | 3.6               | 0.4                |
| Corn price (yuan/kg)             | 2.2                     | 0.3                | 2.2               | 0.3                |
| Commercial feed price (yuan/kg)  | 2.9                     | 0.3                | 3.2               | 0.1                |
| Piglet price (yuan/kg)           | 24.1                    | 6.7                | 32.4              | 9.6                |
| Live hog price (yuan/kg)         | 13.9                    | 2.7                | 15.7              | 2.4                |
| Pork price (yuan/kg)             | 22.3                    | 3.9                | 25.6              | 3.1                |
| Number of observations           |                         | 58                 |                   | 47                 |

**Table 2. Testing for Structural Break Dates**

| Dependent variable | Lag/lead of independent variables | Supremum Wald statistics | Structural break date |
|--------------------|-----------------------------------|--------------------------|-----------------------|
| Sow inventory      | t                                 | 288.2***                 | 05/2014               |
|                    | t-1~t-3                           | 286.6***                 | 10/2013               |
|                    | t-10~t-12                         | 420.7***                 | 12/2013               |
|                    | t+1~t+3                           | 218.4***                 | 04/2014               |
| Hog inventory      | t                                 | 123.5***                 | 02/2015               |
|                    | t-1~t-3                           | 213.9***                 | 12/2014               |
|                    | t-10~t-12                         | 341.4***                 | 02/2015               |
|                    | t+1~t+3                           | 101.1***                 | 01/2015               |

*Note:* Table 2 shows structural break dates estimated by different specifications using the supremum Wald test, as suggested by Andrews (1993). Predictors include prices for corn, soybean meal, commercial feed, hogs, piglets, and pork. A wider search range, when allowed by the degree of freedom, produces similar results. Because estimating a model with a structural break doubles the number of coefficients, it is not feasible to test the specification with all lagged and lead terms.

**Table 3. Hog Inventory Forecasting Performance of SVR Compared to Neural Networks, Random Forest, and OLS**

| Lag/lead structures<br>of predictors | SVR (RBF kernel) |      |      | Neural networks |      |      | Random forest |      |      | OLS   |      |      |
|--------------------------------------|------------------|------|------|-----------------|------|------|---------------|------|------|-------|------|------|
|                                      | RMSE             | NMSE | MAPE | RMSE            | NMSE | MAPE | RMSE          | NMSE | MAPE | RMSE  | NMSE | MAPE |
| Current and leads                    |                  |      |      |                 |      |      |               |      |      |       |      |      |
| t                                    | 11.55            | 1.32 | 2.05 | 11.89           | 1.40 | 2.19 | 12.04         | 1.43 | 1.95 | 7.15  | 0.51 | 1.25 |
| t+1                                  | 13.28            | 1.74 | 2.35 | 15.16           | 2.27 | 2.44 | 11.44         | 1.29 | 2.08 | 6.33  | 0.40 | 1.12 |
| t+2                                  | 13.54            | 1.81 | 2.13 | 11.69           | 1.35 | 2.04 | 10.76         | 1.15 | 1.95 | 5.44  | 0.29 | 0.95 |
| t+3                                  | 8.79             | 0.76 | 1.65 | 11.57           | 1.32 | 2.06 | 12.35         | 1.51 | 2.24 | 6.81  | 0.46 | 1.14 |
| t+1, t+2, t+3                        | 9.31             | 0.86 | 1.71 | 8.47            | 0.71 | 1.55 | 10.71         | 1.14 | 1.88 | 6.14  | 0.37 | 1.03 |
| Recent lags                          |                  |      |      |                 |      |      |               |      |      |       |      |      |
| t-1                                  | 10.68            | 1.13 | 1.90 | 8.99            | 0.80 | 1.72 | 10.66         | 1.12 | 1.80 | 7.69  | 0.59 | 1.28 |
| t-2                                  | 8.03             | 0.64 | 1.39 | 9.50            | 0.89 | 1.65 | 11.55         | 1.32 | 2.07 | 10.81 | 1.16 | 1.73 |
| t-3                                  | 8.75             | 0.76 | 1.67 | 15.18           | 2.28 | 2.72 | 12.50         | 1.55 | 2.27 | 9.41  | 0.88 | 1.52 |
| t-1, t-2, t-3                        | 7.74             | 0.59 | 1.49 | 10.27           | 1.04 | 1.85 | 11.05         | 1.21 | 1.90 | 13.74 | 1.87 | 1.98 |
| Deep lags                            |                  |      |      |                 |      |      |               |      |      |       |      |      |
| t-10                                 | 5.86             | 0.34 | 1.01 | 12.20           | 1.47 | 2.17 | 12.11         | 1.45 | 2.26 | 10.15 | 1.02 | 1.71 |
| t-11                                 | 9.52             | 0.90 | 1.66 | 17.42           | 3.00 | 2.98 | 12.52         | 1.55 | 2.40 | 4.37  | 0.19 | 0.76 |
| t-12                                 | 11.53            | 1.32 | 2.03 | 13.26           | 1.74 | 2.38 | 12.04         | 1.43 | 2.24 | 7.24  | 0.52 | 1.23 |

|   |        |        |        |        |        |        |        |        |        |        |        |       |
|---|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|-------|
| t-10, t-11, t-12                                | 3.89   | 0.15   | 0.75   | 7.84   | 0.61   | 1.23   | 12.61  | 1.57   | 2.28   | 11.61  | 1.33   | 1.82  |
| All and feature selection                       |        |        |        |        |        |        |        |        |        |        |        |       |
| t-3 ~ t+3, t-10 ~ t-12                          | 6.43   | 0.41   | 1.14   | 9.60   | 0.91   | 1.72   | 9.91   | 0.97   | 1.64   | -      | -      | -     |
| Feature selection                               | 4.41   | 0.19   | 0.69   | 5.39   | 0.29   | 0.98   | 10.34  | 1.06   | 1.66   | 11.83  | 1.38   | 2.23  |
| Improvement from<br>feature selection           | -52.1% | -78.8% | -57.7% | -53.7% | -79.7% | -52.0% | -10.8% | -20.9% | -19.9% | 43.8%  | 87.9%  | 65.2% |
| SVR performance<br>relative to other<br>methods |        |        |        | -27.9% | -48.1% | -29.7% | -60.8% | -84.7% | -57.8% | -11.0% | -21.2% | -9.3% |

*Note:* The second-to-last row reports the percentage difference in error measures when using feature selection relative to the average of other specifications. The last row reports the percentage difference of error measures of the best specification in each method relative to the best specification in SVR. OLS regression does not have enough degrees of freedom for the specification with t-3 ~ t+3, t-10 ~ t-12.

**Table 4. Sow Inventory Forecasting Performance of SVR Compared to Neural Networks, Random Forest, and OLS**

| Lag/lead structures<br>of predictors | SVR (RBF kernel) |      |      | Neural networks |       |      | Random forest |      |      | OLS  |       |      |
|--------------------------------------|------------------|------|------|-----------------|-------|------|---------------|------|------|------|-------|------|
|                                      | RMSE             | NMSE | MAPE | RMSE            | NMSE  | MAPE | RMSE          | NMSE | MAPE | RMSE | NMSE  | MAPE |
| Current and leads                    |                  |      |      |                 |       |      |               |      |      |      |       |      |
| t                                    | 0.47             | 2.85 | 0.84 | 0.83            | 8.75  | 1.27 | 0.53          | 3.57 | 0.77 | 0.25 | 0.82  | 0.41 |
| t+1                                  | 0.38             | 1.81 | 0.63 | 0.51            | 3.31  | 0.85 | 0.56          | 4.02 | 0.82 | 0.93 | 11.02 | 1.00 |
| t+2                                  | 0.35             | 1.60 | 0.46 | 0.58            | 4.26  | 0.91 | 0.48          | 2.92 | 0.68 | 0.29 | 1.07  | 0.42 |
| t+3                                  | 0.63             | 5.02 | 0.96 | 0.33            | 1.40  | 0.53 | 0.63          | 5.06 | 0.92 | 0.24 | 0.74  | 0.37 |
| t+1, t+2, t+3                        | 0.25             | 0.79 | 0.37 | 0.25            | 0.78  | 0.39 | 0.41          | 2.10 | 0.60 | 0.98 | 12.22 | 1.15 |
| Recent lags                          |                  |      |      |                 |       |      |               |      |      |      |       |      |
| t-1                                  | 0.41             | 2.18 | 0.67 | 0.84            | 9.05  | 1.25 | 0.67          | 5.77 | 1.04 | 0.43 | 2.39  | 0.69 |
| t-2                                  | 0.58             | 4.25 | 0.83 | 0.73            | 6.73  | 1.23 | 0.46          | 2.64 | 0.72 | 0.64 | 5.14  | 1.05 |
| t-3                                  | 0.55             | 3.80 | 0.84 | 1.11            | 15.62 | 1.41 | 0.43          | 2.34 | 0.71 | 0.61 | 4.67  | 0.85 |
| t-1, t-2, t-3                        | 0.52             | 3.39 | 0.75 | 0.60            | 4.60  | 1.03 | 0.53          | 3.53 | 0.93 | 0.66 | 5.53  | 1.08 |
| Deep lags                            |                  |      |      |                 |       |      |               |      |      |      |       |      |
| t-10                                 | 0.28             | 0.97 | 0.46 | 0.79            | 7.98  | 0.91 | 0.49          | 3.03 | 0.73 | 0.19 | 0.46  | 0.30 |
| t-11                                 | 0.32             | 1.27 | 0.46 | 1.14            | 16.57 | 1.49 | 0.47          | 2.83 | 0.77 | 0.28 | 1.01  | 0.37 |
| t-12                                 | 0.26             | 0.86 | 0.41 | 0.87            | 9.65  | 1.44 | 0.43          | 2.36 | 0.61 | 0.20 | 0.51  | 0.29 |

|   |        |        |        |        |        |        |        |        |        |        |        |        |
|---|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| t-10, t-11, t-12                                | 0.26   | 0.86   | 0.41   | 0.70   | 6.21   | 1.04   | 0.52   | 3.42   | 0.87   | 0.44   | 2.49   | 0.66   |
| All and feature selection                       |        |        |        |        |        |        |        |        |        |        |        |        |
| t-3 ~ t+3, t-10 ~ t-12                          | 0.22   | 0.62   | 0.34   | 0.87   | 9.76   | 1.16   | 0.43   | 2.40   | 0.73   | -      | -      | -      |
| Feature selection                               | 0.17   | 0.35   | 0.25   | 0.56   | 3.97   | 0.75   | 0.39   | 1.95   | 0.64   | 1.12   | 15.88  | 1.69   |
| Improvement from<br>feature selection           | -57.4% | -83.8% | -58.4% | -23.0% | -47.0% | -29.3% | -22.1% | -40.6% | -17.3% | 136.5% | 329.5% | 154.5% |
| SVR performance<br>relative to other<br>methods |        |        |        | -32.8% | -55.1% | -35.6% | -57.5% | -82.0% | -58.2% | -12.2% | -22.9% | -14.3% |

*Note:* The second-to-last row reports the percentage difference in error measures when using feature selection relative to the average of other specifications. The last row reports the percentage difference of error measures of the best specification in each method relative to the best specification in SVR. OLS regression does not have enough degrees of freedom for the specification with t-3 ~ t+3, t-10 ~ t-12.