경제학석사학위논문

Data-driven Agricultural Market Prediction

- Using Data from Hog-farm Management Information System -

정보시스템 데이터를 활용한
시장 예측에 관한 연구

2019년 2월

서울대학교 대학원

농경제사회학부 지역정보전공

황 예 슬

Abstract

Data-driven Agricultural Market Prediction
- Using Data from Hog-farm Management Information System -

Yeseul Hwang

Regional Information Major
Graduate School of Agricultural Economics
and Rural Development

Seoul National University

In the information age, many economic problems have been solved using accumulated data from information systems. This study tries to relieve two controversial issues in market prediction using such data. First, the controversies often exist over the theoretical model, Rational Expectation Hypothesis (REH) model, and the empirical model, the time-series model. REH has been questioned for its overly restrictive assumptions and its poor forecasting performance. Second, this study tries to verify their better predictive power of machine-learning method over time-series method in the case with the agent-level data. The machine-learning methods in this study adopt disaggregated farm-level data to improve flexibility, which has not been attempted in previous studies. They are compared with the conventional time-series methods. Consequently, the explanatory and predictive power of the REH model was found to be worse than the model with farm-level data. Also, the time-series model dominates the REH model in supply forecasting,

which is consistent with previous studies. Furthermore, machine-learning methods using disaggregated data performed better than the time-series models, when forecasting the far future.

Keywords : Rational Expectation Hypothesis, Support Vector Regression, Artificial Neural Network, Supply Response Analysis, Price Forecasting, Farm-level Data

Student Number : 2016-21492

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

The world is usually in disequilibrium and variance. Forecasting human behavior in social science is particularly challenging. Price forecasting in the market is a complicated task, especially for the agricultural product market, as diverse factors, including environmental risks, affect the market. Technology improvement has made the accumulation of data in various domains possible. Also, people can access this data more easily than ever before. Many attempts have been made to improve the quality of human life by using such data in business and research in various domain (Choe & Lee, 2010; Fornaro, 2016; Marcjasz, Uniejewski, & Weron, 2018; Taylor, McSharry, & Buizza, 2009; Wang, Rothschild, Goel, & Gelman, 2015; Wolfert, Ge, Verdouw, & Bogaardt, 2017; Ye, Zyren, & Shore, 2005).

This study aims to address two controversial issues in the literature relating to market prediction. The rational expectation hypothesis (REH) will be analyzed in terms of its explanatory and predictive power (Shmueli, 2010). Also, this study will explore whether machine-learning methods predict markets better than other methods. These become testable due to the availability of data from various information systems.

First, many researchers have questioned whether REH is

empirically observable (e.g., the international economic crisis in 2008; (Hommes, 2011). Even though REH is a fundamental theory of decision-making for humans and is widely used in economic research, it has been criticized for its overly restrictive assumptions (Guidolin & Thornton, 2018; Hendry & Muellbauer, 2018; Sims, 1986). Therefore, this study contributes to the REH literature by comparing a model based on REH with a farm-level data model and a time-series model.

Second, while many studies have compared machine-learning methods to time-series methods, the results of a few studies show that the machine-learning method does not produce superior predictions (Adya & Collopy, 1998; Bloznelis, 2018; Chakraborty, Mehrotra, Mohan, & Ranka, 1992; Hamm & Brorsen, 1997; Shahwan & Odening, 2007; Yun, Lee, & Yang, 2016). However, most of these studies have traditionally used aggregated data only. Because the machine-learning method relaxes many assumptions of the statistical method, it is less likely to perform better with aggregated data. Therefore, this study tries to clarify the usefulness of the machine-learning methods using disaggregated data, compared with the time-series methods.

The paper is organized as follows. Chapter 2 provides a brief explanation of the hog market and life cycle to help understand industry structure and data used in this study. Then, it reviews the studies related to market prediction. The theoretical background focuses on REH, time-

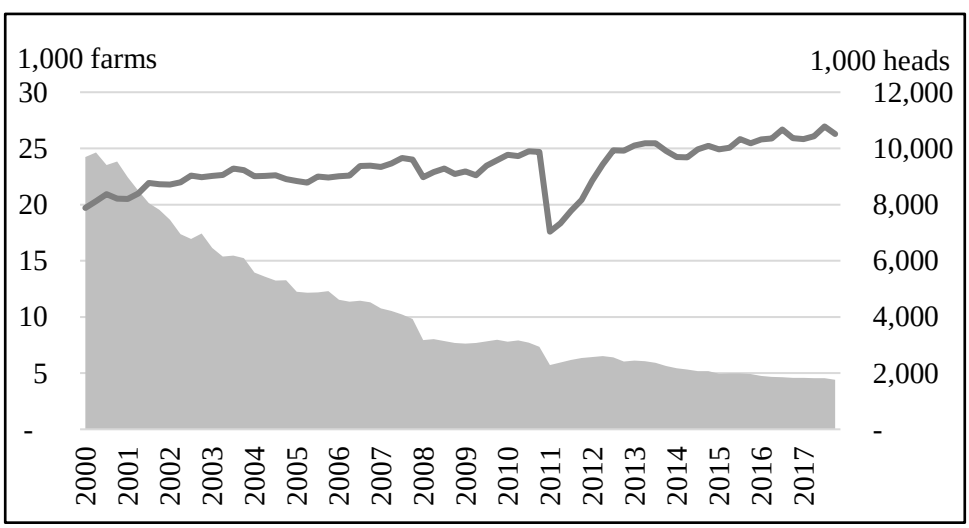
series models, and machine-learning methods, which are controversial issues in market prediction. Chapter 3. In chapter 4, models adopted in this study are explained in the following order: model based on REH, model with the aggregated farm-level data, time-series models, and machine-learning models. In chapter 5, three models are tested and compared each other to answer the two main research questions raised in this study. To test the REH, the model based on REH, the model with the aggregated farm-level data, and the time-series models are compared. Then, the time-series models and machine-learning models are compared to clarify the usefulness of machine-learning methods using disaggregated data. Chapter 6 provides the conclusions of this study.

Chapter 2. Background of Study

2.1. Industry Trend of Pork Market

To help understand the study, this chapter summarizes some background knowledge regarding the hog market and the hog lifecycle. The size of the hog market in South Korea has grown over the past five years. The GDP of the hog industry was almost 7.7 billion KRW in 2016, an increase of 26.3% compared to 2012. Also, the hog industry has grown in terms of its contribution to GDP, in proportion not only to the agriculture industry but also to the livestock industry.

The average scale of each hog farm is higher than before. From 2000 to 2017, over 80% of the hog farmers shut down (Figure 1).

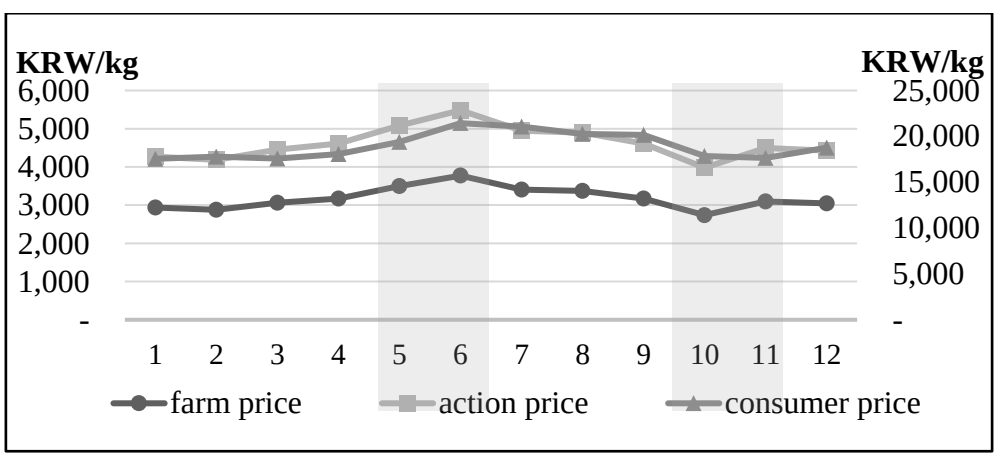


Source: Statistics Korea

Figure 1 The Number of Hogs and Farms in South Korea

However, the number of fed hogs has increased[①]. In the fourth quarter of 2017, 4,406 farms raised 10.5 million hogs. There is a clear negative relationship between the number of hogs and price in the auction. More importantly, hog price has strong seasonality as shown in Figure 2. The conception rate is higher in winter than summer, because sows are typically under more stress during summer. The higher conception rate in winter results in increased supply and reduced prices in October and November. Similarly, the lower conception rate in summer results in reduced supply and higher prices in May and June (Iida & Koketsu, 2013; Jung & Kim, 2011).

The growth of the industry has amplified the need for accurate forecasting of hog supply and price. Also, the large size of each farm has



Source: Korea Institute for Animal Products Quality Evaluation

Figure 2 Seasonal Trend of Hog Price (average of 2003 ~ 2017)

① In 2011, severe foot-and-mouth disease proliferated across the country, causing massive hog slaughter.

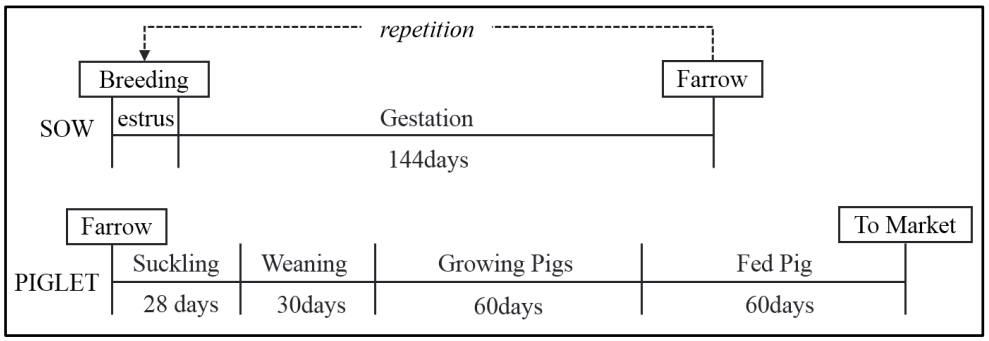


Figure 3 Hog Production Process

made precise forecasting increasingly important. The unpredictability in hog supply and price poses a challenge to the stable management of farms in the short term and the secure development of the industry in the long term (Kim, 1998).

It is important to know the fed-pig production process, because it is highly related to the lag selection for the study. When a sow succeeds in gestation after breeding, it takes 144 days to farrow. Once born, the piglets go through the suckling stage for 28 days. After that, they become weaning piglets. After 10 months from gestation, 6 months from farrow, and 5 months from weaning, fed pigs are sold in the market (Figure 3).

Farmers maximize their profit under fixed assets (the spatial size of the farm). Feed costs make up for more than half of the total input cost required to raise a pig. If the feed cost is high, managers can sell hogs earlier by taking the risk of earning less money. If they sell hogs too early, they get paid less because the market price is linked to the weight of each hog.

2.2. Theoretical Background

To explain the decision-making process of humans, economists assume that people make decisions based on their expectation on the market. Their expectation formation is important in predicting market changes. The two main expectation hypotheses are adaptive expectation hypothesis and rational expectation hypothesis.

Nerlove (1958) proposed a dynamic supply response analysis using the adaptive expectation hypothesis. Adaptive expectation is a process through which people form expectations of market supply and price based on past information. In the 1960s, most of the research predicting commodity markets took advantage of this hypothesis (Goodwin & Sheffrin, 1982; Holt & Johnson, 1988). However, adaptive expectation assuming the fixed reflection rate of expectation error could not avoid systemic errors.

Muth (1961) devised the rational expectation hypothesis to fix these systemic errors and the concept has been actively adopted after Lucas (1976) applied it to macroeconomics. Compared with the adaptive hypothesis, REH does not have a particular form of expectation. For instance, the expected price can be expressed as conditional expectation on all available information at the time and shown in equation 2.1.

$$p_t^e = E(p_t|\Omega_{t-1}) \tag{2.1}$$

where Ω_{t-1} is an information set that contains all information available in t-1. The rational expectation process is the functional relationship between the future price and information set, embracing the relationship between variables. Under the REH, all individuals are the same and forecast rationally, using available information. The systemic error of rational expectation appears only when one does not make use of the complete information set, or if one uses the information set in an invalid way. Since the 1980s, many studies have presumed rational expectation to predict the market (Antonovitz & Green, 1990; Goodwin & Sheffrin, 1982; Klein, 2000; Runkle, 1991; Wallis, 1980). Consequently, REH became a crucial part of economic theory (Hommes, 2011).

However, REH has been criticized for its strong assumptions while ignoring heterogeneity. Indeed, individuals are heterogeneous in terms of information accessibility and information processing abilities (Haltiwanger & Waldman, 1985; Lovell, 1986). Furthermore, the models based on REH are often empirically unobservable; the international economic crisis in 2008, which led to the decline of worldwide financial markets by almost 50% between October 2008 and March 2009, is hard to explain using rational behavior; (Hommes, 2011). These drawbacks pose a risk to the utility of REH.

Sims (1986) criticized the economic models that assume rational

expectation, because they often forecast the future poorly due to the simplified assumptions in REH. He insisted that we know too little about the nature of dynamic economic behavior, leading to endogeneity problems in the model. Such models are more likely to have illogical or non-optimal relationships with the data and their results can be incorrect and unreliable for practical use. He alternatively affirmed the use of time-series analysis and vector autoregression (VAR). This skepticism has continued till date. Guidolin and Thornton (2018) compared different forecasting models, in which the expectation hypothesis is assumed. However, their result was worse than the random forest benchmark model. They concluded that the assumptions about the market participants' ability to predict the future are not appropriate. Hendry and Muellbauer (2018) insisted that the world is usually non-stationary and often unanticipated, for REH to forecast effectively. Therefore, they argued that the theory-driven approaches must be integrated with the data-driven approaches to overcome the weakness of a simplistic and unrealistic theory.

Since then, many predictive modeling studies in agricultural economics have adopted time-series models, such as autoregressive integrated moving average (ARIMA) or VAR (Antonovitz & Green, 1990; Brandt & Bessler, 1981; Holt & Johnson, 1988). Time-series analysis has attracted attention for its less restrictive assumptions and

high forecasting power, compared to REH (Choi, 2016; Sims, 1986).

Prediction models based on REH and time series both use aggregated data. However, more disaggregated data resulted from individual decision has recently become available owing to a reduction in the cost of data processing and storage, and an improvement in communication technology. Consequently, there is a movement toward the use of individual-level data generated from information systems to solve problems, which could not be possible before (Bragoli, 2017; Fornaro, 2016; Lessmann & Voß, 2017; Powell et al., 2017). The benefit of using disaggregated data is that the assumptions about individual behavior are no longer needed.

Market prediction using disaggregated data is related to machine-learning methods. Machine learning is a branch of artificial intelligence that synthesizes the underlying relationships systematically and among data. It is applied on engineering and natural science at first and now expanded to data from other fields (Awad & Khanna, 2015). Market prediction using machine-learning methods has been increased in domains where massive data collection is feasible, including finance (Kaastra & Boyd, 1996; Shahwan & Odening, 2007), agricultural product (Harchaoui & Janssen, 2018), electricity (Marcjasz et al., 2018) and even enhancement of aggregated data. However, the number of studies is insufficient.

Machine-learning methods look promising because they are much more flexible than regression models including time-series models. Machine-learning methods are expected to relex many priori assumptions of regression models. Additionally, they do impose less restrictions on variables or data. So, they find better relationship between data, especially when the relationship is non-linear.

This study tried to compare diverse models to answer two main questions related to the literature. First, the model based on REH will be analyzed. Even though various studies have asserted the weak forecasting power of the REH, this has not been proven with data from diverse backgrounds due to the lack of access to the data (Chow, 1989; Figlewski & Wachtel, 1981; Lovell, 1986; Weizsäcker, 2010). As disaggregated data has been accumulated from information system, the data more reflective on the future information set Ω_t become available. As a result, a model based on REH can be empirically tested of its forecasting power by comparing it with a model using such micro-level data and a time-series model (Shmueli, 2010).

Second, this study tries to validate the usefulness of machine learning in market prediction. Despite the flexibility of machine-learning methods, some studies are reluctant to use machine-learning techniques, because their forecasting performance has not been outstanding when compared with time-series analysis (Adya & Collopy, 1998; Bloznelis,

2018; Chakraborty et al., 1992; Hamm & Brorsen, 1997; Shahwan & Odening, 2007). Yet, it is still early to criticize the machine-learning methods for its predictability. These researches have used aggregated data which has limits in the amount of the data. Because the machine-learning method works better with the large data set, it is less likely to perform better with limited amount of aggregated data which ignores difference in individual decision making. Therefore, this study will clarify the usefulness of the machine learning method using disaggregated data. To address these two research questions, this study uses hog farm-level data collected by an information system.

Chapter 3. Methods

This comprehensive study is composed of two related steps as shown in Table 1. In the first step, the model based on REH will be compared with the model using aggregated farm-level data and a time-series model in terms of their explanatory and predictive powers. By doing this, the model based on REH can be evaluated based on the discussions from previous studies. In the second step, machine-learning methods with disaggregated data will be compared with the time-series methods with aggregated data.

	Compared Models	Evaluation Criteria	Aim
Step 1	Model based on REH vs. Model with the Aggregated Farm-level Data	R^2	Empirically test REH from a comprehensive perspective
	Model based on REH vs. Model with the Aggregated Farm-level Data vs. Time Series Model (VAR)	Forecasting Error	
Step 2	Time Series Models (SARIMA & SVAR) with Aggregated Data vs. Machine Learning Methods (SVR vs. ANN) with Disaggregated Data	Forecasting Error	Validate the usefulness of the machine learning in market prediction.

Step 1 uses only monthly data, but Step 2 uses monthly and weekly data

Table 1 Summary of the Compared Models

3.1. Supply Response Model based on REH

To test the explanatory and predictive power of the REH model, the hog supply response model will be used. Estimating a supply response model is an important task in agricultural economics, because the price of agricultural products is often inelastic, making supply the most important factor affecting price. Accordingly, many studies have tried to identify the hog supply model (Arzac & Wilkinson, 1979; Holt & Johnson, 1988; Leuthold, MacCormick, Schmitz, & Watts, 1970; Meilke, Zwart, & Martin, 1974; Prescott & Stengos, 1987; Reutlinger, 1966; Tryfos, 1974).

In this study, the theoretical model of Antonovitz and Green (1990) is adopted with some modification. The supply and demand function suggested by them is shown below.

$$Q_t = a + \sum_{i=1}^{5} b_i S_{it} + cFP_{t-r} + dP_t^e + \varepsilon_t \qquad (4.1)$$

$$P_t = e + fQ_t + gY_t + \sum_{j} h_j p_t^j + u_t \qquad (4.2)$$

The variables are summarized in Table 2 and ε_t and u_t are error terms for each model.

By satisfying the equilibrium condition, the optimal P_t and Q_t are derived. The detailed derivation can be found in Antonovitz and Green (1990) article.

Name of Variables	Definitions
$\boldsymbol{Q_t}$	Hog supply at period t
$\boldsymbol{FP_t}$	Average feed price from period t-4 to period t
$\boldsymbol{Y_t^e}$	Expected income of period t
$\boldsymbol{p_t^{be}}$	Expected beef price of period t
$\boldsymbol{p_t^{ce}}$	Expected chicken price of period t
$\boldsymbol{W_{t-5}}$	The number of piglets who just started weaning at period t-4
$\boldsymbol{S_{it}}$	Seasonal Dummy; spring($i = 1$), summer($i = 2$) and fall($i = 3$)

This table defines all the variables covered in chapter 4

Table 2 Variable Definitions

$$P_t^* = e + fQ_t + gY_t + \sum_j h_j p_t^j + u_t \tag{4.3}$$

$$Q_t^* = \frac{a + de}{1 - fd} + \frac{\sum_{i=1}^{5} b_i S_{it}}{1 - fd} + \frac{c}{1 - fd} FP_{t-r} + \frac{dg}{1 - fd} Y_t^e + \frac{d}{1 - fd} \sum_{j=1}^{2} h_j p_t^{je} + \varepsilon_t \tag{4.4}$$

The supply function can be expressed in the econometric form as below:

$$Q_t = \beta_0 + \sum_{i=1}^{5} \beta_i S_{it} + \beta_6 FP_t + \beta_7 Y_t^e + \beta_8 p_t^{be} + \beta_9 p_t^{ce} + \varepsilon_t \tag{4.5}$$

where $Y_t^e = E(Y_t|\Omega_{t-1}) = Y_t$,

and $p_t^{je} = E(p_t^j|\Omega_{t-1}) = p_t^j$

3.2. Model with the Aggregated Farm-level Data

The supply response model will be compared to the model with farm-level data. The model is constructed based on the lifecycle of the hogs. This model becomes testable due to the data accumulated through the information system. As shown in equation 4.6, the model uses only the number of weaned piglets as an independent variable, other than the seasonal dummy variable. Because weaned piglets become hogs for sale biologically after five months, the number of weaned piglets is expected to be highly correlated with the hog supply. Whether the REH model will dominate this model is the key point.

$$Q_t = \beta_0 + \sum_{i=1}^{3} \beta_i S_{it} + \beta_6 W_{t-5} + \epsilon_t \qquad (4.6)$$

In addition to the simple model, a combination model will be also compared, which contains all the economic variables and the number of weaned piglets, as shown in equation 4.7. Because weaned piglets are strongly related to the supply, the weaned-piglet variable is expected to take away some significance of the other variables.

$$Q_t = \beta_0 + \sum_{i=1}^{5} \beta_i S_{it} + \beta_6 FP_t + \beta_7 Y_t^e + \beta_8 p_t^{be} + \beta_9 p_t^{ce} + \beta_{10} W_{t-5} + \epsilon_t \qquad (4.7)$$

3.3. Time-series Models

The time-series model is employed to compare the forecasting results, not only with the model based on REH but also with the machine-learning methods with disaggregated data. In this section, a brief summary of two time-series methods is provided, based on the book "Forecasting: Principles and Practice" (Rob J Hyndman & Athanasopoulos, 2018).

To use the time-series analysis, time-series data have to be stationary. In other words, it should not have a unit root. Stationary time-series data must satisfy the assumption of constant mean and variance. In this study, both the Phillips-Perron test and augmented Dickey-Fuller test are employed; *urpptest* and *urdfTest* in *fUnitRoots* package of R are used. If unit root exists, the time series can be made stationary through differencing.

The first method is seasonal ARIMA, or SARIMA. It is an extension of ARIMA. The time series follows ARIMA(*p, d, q*), where *p* is the order of the autoregressive part, *d* is a degree of first differencing involved, and *q* is the order of the moving average part. ARIMA(*p, d, q*) can be expressed as

$$\phi(B)(1-B)^d Y_t = \theta(B)\varepsilon_t \tag{4.8}$$

where $\phi(B) = 1 - \phi_1 \mathrm{B} - \phi_2 \mathrm{B}^2 - \cdots - \phi_p \mathrm{B}^p$ and $\theta(B) = 1 +$

$\theta_1 \mathrm{B} + \theta_2 \mathrm{B}^2 + \cdots + \theta_q \mathrm{B}^q$

Autocorrelation function and partial autocorrelation function correlogram are used to determine the orders of the models p and q, respectively. In our case, it is consistent with the order selection for minimizing the Akaike information criterion. Furthermore, when using monthly data, seasonality is adjusted by taking seasonal differentiation into consideration. By contrast, when using weekly data, Fourier terms are adopted to deal with seasonality.

The second method is seasonal VAR. VAR generalizes the univariate restriction of ARIMA. Under VAR, Y_t is now a vector of endogenous variables $Y_t \in \mathbb{R}^n$and $\varepsilon_t \in \mathbb{R}^n$, where n is the number of endogenous variables. Also, ϕ_p is an $\mathrm{n} \times \mathrm{n}$ matrix and θ_q is an $\mathrm{n} \times \mathrm{m}$ matrix of coefficients. The orders are chosen by minimizing the AIC. Seasonality is treated as above.

Both ARIMA and VAR take only stationary time series. However, even though the data is non-stationary, it can be analyzed without solving the problem if there is cointegration. In this study, the Johansen cointegration test was conducted to test whether there is cointegration; *ca.jo* function in *urca* package of R was used. If the tests indicate that there is cointegration, the vector error correction model (VECM) is utilized instead of VAR.

3.4. Machine-learning Methods

In this section, two machine-learning methods, support vector regression (SVR) and artificial neural network (ANN), are summarized, based on the book "Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers," written by Awad and Khanna (2015). SVR and ANN are widely used to solve the regression problem. They are much more flexible compared with multiple regression by least squared estimation. This study uses these methods to deal with the disaggregated data.

SVR is a generalized model of support vector machine in that it is applicable to regression problems. Unlike ANN, it always returns the same optimal hyperplane parameter, because it solves the convex optimization problem analytically.

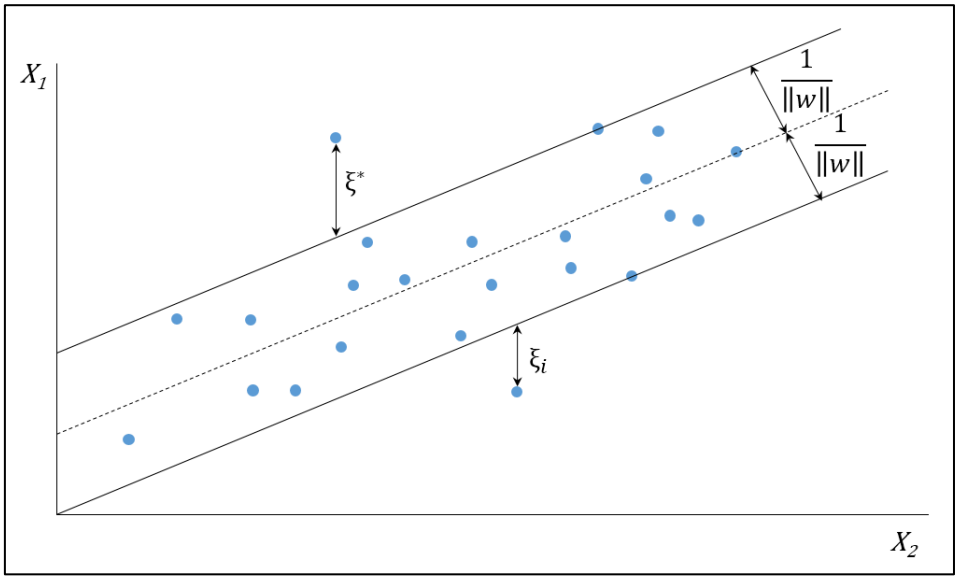
$$y = f(x) = \begin{bmatrix} w \\ b \end{bmatrix}^t \begin{bmatrix} x \\ 1 \end{bmatrix} = w^t x + b \; x, w \in \mathbb{R}^{M+1} \qquad (4.9)$$

The optimization problem of SVR is to find the narrowest tube centering, with $w^t x + b = 0$, while minimizing the sum of prediction errors. This problem can be represented as given below.

$$\min_w \frac{1}{2} \|w\|^2 \qquad (4.10)$$

subject to the following constraint:

$$y_i(w_i^T x + b) \geq 1, \qquad i = 1, 2, \dots, N$$



Source: Awad and Khanna (2015) (modified)

Figure 4 Support Vector Regression

Slack variables ξ_i can be added to the SVR model. By introducing ξ_i to the objective function, the model allows some observations outside the tube. C is the regulation term, which should be determined by the researchers. A higher C is less likely to accept violation, because it results in a narrower tube.

$$\min \frac{1}{2}\|w\|^2 + \mathrm{C}\sum_{i=1}^{N} \xi_i + \xi^* \qquad (4.11)$$

subject to the following constraints:

$$y_i - w^T x_i \leq \varepsilon + \xi^*, \qquad i = 1, 2, \ldots, N$$

$$w^T x_i - y_i \leq \varepsilon + \xi^*, \qquad i = 1, 2, \ldots, N$$

$$\xi_i, \xi^* \geq 0, \qquad i = 1, 2, \ldots, N$$

Kernel is used to transform the data into a higher dimensional space, the kernel space, to accurately regress with a linear hyperplane. There are

various kernel functions, but in this study Gaussian radial basis function is used. It is known as a general –purpose kernel that is mostly applied in the absence of prior knowledge. It can be described as given below.

$$K(x,u) = exp\left(-\frac{\|x-u\|^2}{\sigma^2}\right) \tag{4.12}$$

σ affects the distribution of the data in the kernel space. The data are more widely scattered as σ decreases. Now, the SVR objective function must satisfy the Karush–Kuhn–Tucker conditions, and can be represented as given below.

$$\min \frac{1}{2}\|w\|^2 + \mathrm{C}\sum_{i=1}^{N} \xi_i + \xi^* \tag{4.13}$$

subject to the constraints:

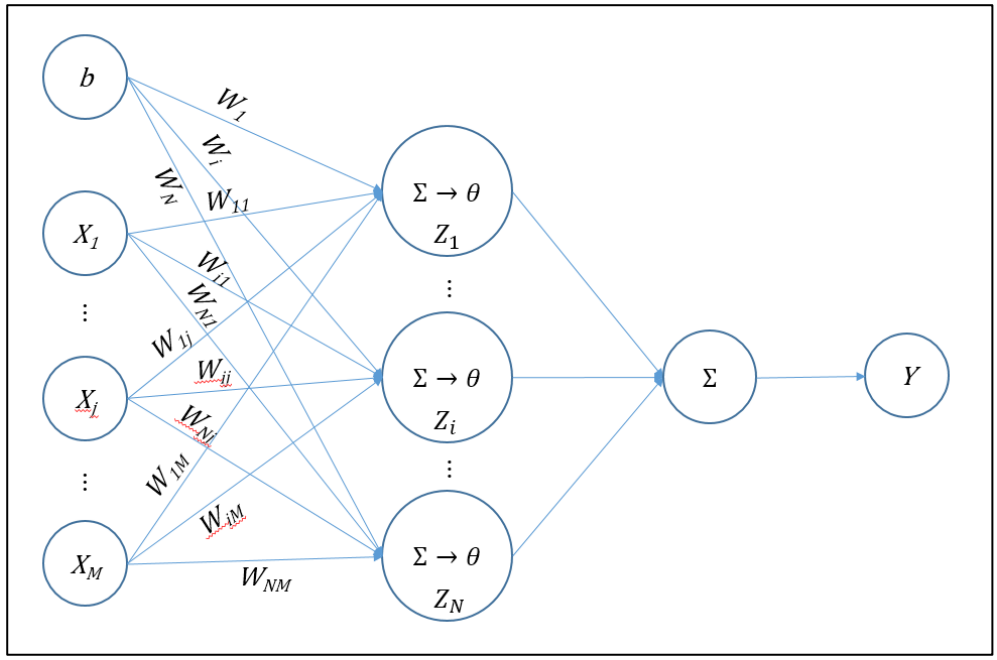
$$y_i - w^T\varphi(x_i) \le \varepsilon + \xi^*, \qquad i = 1, 2, \dots, N$$

$$w^T\varphi(x_i) - y_i \le \varepsilon + \xi^*, \qquad i = 1, 2, \dots, N$$

$$\xi_i, \xi^* \ge 0, \qquad i = 1, 2, \dots, N$$

The R package “e1071” is used to implement the SVR model.

ANN is another powerful machine-learning method to solve the regression problem. ANN is simply an association of layers of neurons, each with its own weight matrix, bias vector, and output vector. If an input vector is constituted of N inputs and a layer of M neurons, W_{ij} represents the weight of the connection of the j^{th} input to the i^{th} neuron of the layer; Y and b are, respectively, the output of and the bias. Each



Source: Awad and Khanna (2015) (modified)

Figure 5 Artificial Neural Network

neuron consists of a summation function and an activation function. Figure 5 shows the illustration of the artificial neural network. There are various types of activation functions, but based on previous studies, a sigmoid function is used in this study (Bloznelis, 2018).

$$\theta(a) = \frac{1}{1 + e^{-a}} \qquad (4.14)$$

Z_i, the output from node I, is

$$Z_i = \frac{1}{1 + \exp(W_i + W_{i1}X_1 + W_{i2}X_2 + \cdots + W_{iM}X_M)} \qquad (4.15)$$

Therefore, predicted Y becomes

$$Y = \nu_0 + \nu_1 Z_1 + \nu_2 Z_2 + \cdots + \nu_N Z_N \qquad (4.16)$$

Then the weights and biases are updated with

$$W_{ij}^{k}(t+1) = W_{ij}^{k}(t) - \alpha \frac{\partial E}{\partial W_{ij}^{k}} \tag{4.17}$$

$$b_{i}^{k}(t+1) = b_{i}^{k}(t) - \alpha \frac{\partial E}{\partial b_{i}^{k}} \tag{4.18}$$

where $E = \frac{1}{2}\sum_{i=1}^{P}(Y_i - T_i)^2$, and k is the order number of the hidden layers. The update stops when the error increases.

This does not have a unique mathematical optimization, because the optimal number of layers and neurons for best performance is analytically undetermined. Therefore, a number of trials and errors are needed to attain a more accurate model. Additionally, researchers using ANN have to be cautious, because this method is based on calculation, not estimation. There is always an overfitting problem, because it is a data-driven modeling approach; it calculates weights based on the existing data.

3.5 Seasonality Issues

The problems with seasonality are detailed in two parts. The first problem is that the machine-learning method does not recognize time-series data. Even though the seasonality of monthly data is obvious, dealing with seasonality is difficult. Because the machine learning

method considers the data to be independent, the relative price approach has been widely employed to adjust for seasonality (Peel & Meyer, 2002). A variable can be decomposed into two levels to construct the model that focuses on the short-term market volatility (Ye et al., 2005). One is the normal level, which contains past changes and trends, depending on seasonality. The other is the relative level, which is the gap between the original observation and the normal trend. It shows the short-term market variance. This relative price approach is not new, as studies in macroeconomics, especially those covering relationships between price and inflation, have used relative price (Parsley, 1996). Aggregated price tends to mask the degree of relative price variability (Danziger, 1987). Therefore, the relative price can clarify the effects of the model. Accordingly, when forecasting with monthly data by the machine learning method, the relative price level is used instead of the actual price to capture seasonality.

$$RP_{yt} = P_{yt} - P_t^* \tag{4.8}$$

RP_{yt} is the relative price at the t^{th} week of year y, where $t \in \{1, 2, \cdots, 52\}$. P_{yt} is the actual price at the t^{th} week of year y. P_t^* is the normal price level, which is the mean value of the price at week t of all years.

The second problem is that seasonality is obscure in weekly data. Hyndman (2010) suggests that an unsmooth pattern of long

	Time Series Analysis	Machine Learning Technique
Monthly Data	seasonal differentiation and use seasonal dummy variable	use relative price instead of original price value and use seasonal dummy variable
Weekly Data	adopt Fourier-term	adopt Fourier-term

Table 3 Dealing with Seasonality

seasonal periods can be dealt with using the Fourier series approach. The example is shown below.

$$y_t = a + \sum_{k=1}^{K} \left[\alpha_k \sin\left(\frac{2\pi kt}{n}\right) + \beta_k \cos\left(\frac{2\pi kt}{n}\right) \right] + M_t \qquad (4.9)$$

where n is the total number of seasonal periods and t is one of them. M_t indicates an original model, such as the ARIMA process. The value of *K* can be chosen by minimizing the AIC. This technique has been adopted by studies that use weekly or daily data (Bruhns, Deurveilher, & Roy, 2005; Fulford, Rayco-Solon, & Prentice, 2006; Rayco-Solon, Fulford, & Prentice, 2005; Taylor et al., 2009). Table 5 summarizes how we deal with seasonality in this study.

3.6. Assessment Criteria of Forecasting Error

A variety of error assessment measures are used in this study. Each criterion has its own characteristic mean absolute error (MAE) and root mean square error (RMSE), which are commonly used among scale-dependent measures. MAE is easy to interpret and the nature of its loss

function is weaker (Nikolopoulos, Goodwin, Patelis, & Assimakopoulos, 2007). By contrast, RMSE is theoretically related to statistical modeling. However, it is more sensitive to outliers than MAE (Rob J. Hyndman & Koehler, 2006).

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - f(x_i)|}{n}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - f(x_i))^2}{n}}$$

Because these two methods are scale dependent, they can be useful only in comparing different methods applied to the same set of data. So, scale-independent measures should be used to compare forecast performance across data sets. Mean absolute percentage error (MAPE) and root mean square percentage error (RMSPE) are two frequently used measures.

$$\mathrm{MAPE} = \frac{\sum_{i=1}^{n} |\, y_i - f(x_i) / y_i \,|}{n}$$

$$\mathrm{RMSPE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - f(x_i) / y_i)^2}{n}}$$

The earlier measures care only about the size of the gap between actual value and the predicted value. However, even if the size of the forecasts errors is small, it might not forecast the direction well. Forecasting direction of change is also important with time-series data

(Witt & Witt, 1989). Therefore, in addition to the measures above, turning point error (TPE) is employed in this study. TPE occurs when the forecast misses the actual direction of change.

$$\text{TPE} = \frac{\text{number}\left(\text{sign}(f(x_t) - y_{t-1}) \neq \text{sign}(y_t - y_{t-1})\right)}{n}$$

Chapter 4. Analysis and Result

4.1. Data

The description of the data will be provided both on monthly and weekly basis, and it is summarized in Tables 4 and 5. The time period under investigation is from January 2012 to March 2018 (75 months, 315 weeks). First, data on hog supply and price are provided by the Korea Institute for Animal Products Quality Evaluation through eKAPEPIA (http://www.ekapepia.com). The supply of the hog, which does not imply demand, is calculated from the number of heads slaughtered. The price

Name of Variables	Average	S.D	Skewness	Kurtosis	Min.	Median	Max.
P_t	4162.31	628.44	-0.288	2.662	2669	4185	5374
Q_t	1,324,354	140,686	0.162	2.660	725,862	1,200,926	1,601,523
FP_t	595.72	38.55	0.010	1.494	320.5	550.7	658.5
Y_t	3,498,854	135,362	-0.116	2.356	2,114,000	2,966,000	3,767,000
p_t^b	14,224	1,900	0.167	1.885	7,183	12,111	17,587
p_t^c	1,951.48	273.24	0.218	2.524	799.70	1,816.60	2,664.60
W_t	299.17	48.77	0.232	2.560	43.12	210.08	407.04

Table 4 Descriptive Statistics of Monthly Data

Name of Variables	Average	S.D	Skewness	Kurtosis	Min.	Median	Max.
P_t	4,399	711.28	-0.170	2.641	2,731	4,404	5,984
Q_t	305,153	58,054	-1.424	8.242	256	312,315	484,669
FP_{t-r}	595.70	38.28	0.010	1.488	537	604	659
p_t^b	868.10	452.73	-0.053	1.885	24	862	1,633
p_t^c	1,951	304.52	0.129	2.483	1,223	1,927	2,690
W_t	79.95	33.51	4.104	20.195	44.67	73.87	251.25
TO_t	7.45	1.42	-0.28	2.47	4.26	7.61	8.32

Table 5 Descriptive Statistics of Weekly Data

of chicken is available from Korea Broiler Council (http://www.chicken.or.kr). These data are generated on a daily basis and aggregated to weekly or monthly data for this study. The average hog supply is 1,324,354 heads per month. The average beef price is 14,224 KRW. The mean price of chicken is approximately 1,951 KRW.

On the 20th of every month, the Ministry of Agriculture, Food and, Rural Affairs announces the average price of compound feed for swine, based on producer sale price for the month. In this study, the price of compound feed for swine is used but it will be referred to as feed price for convenience. Average feed price is about 596 KRW/kg. Nominal disposable income per household is used as the income variable in this study. The average monthly income is about 3.49 million KRW.

The number of piglets that have just started weaning is derived using data from “Pig Plan,” an information system for swine farms. This

Year	Number of New users	Cumulative Number of Total Users
2010	54	280
2011	87	367
2012	83	450
2013	55	505
2014	55	560
2015	56	616
2016	97	713
2017	85	798

Table 6 Number of "Pig Plan" Users by Year

Breeding heads	"Pig Plan" user	Population	Ratio	Breeding heads of "Pig Plan" user	Breeding heads of the population	Ratio
Less than 1,000	61	1,656	3.68%	34,883	819,256	4.26%
1,000~5,000	340	2,425	14.02%	1,106,707	5,470,156	20.23%
5,000~10,000	169	321	52.69%	1,450,711	2,137,623	67.87%
10,000 이상	114	117	97.44%	2,067,200	2,086,717	99.06%
Total	684	4,518	15.14%	4,659,501	10,513,752	44.32%

- Ratio shows the ratio of the number of pig users to the total population

Table 7 "Pig Plan" Users and Population by Breeding Heads

system was developed by Ezfarm (www.ezfarm.co.kr) in 1999 and became the only actively adopted information system by swine farms (Choe & Lee, 2010). In "Pig Plan," each farmer and their sows get distinct ids and then the events of the farming, such as gestation, farrowing, and weaning of piglets are recorded. As shown in Table 6, the number of total users keeps increasing; it almost tripled in 2017, when compared to 2010. The "Pig Plan" data covered 15.1% of the hog farms

and 44.3% of hogs produced yearly. Most of the big size farmers employ the system, but only a limited number of small farmers use it (Table 7).

Therefore, “Pig Plan” data cannot be considered as a representative sample. Because the data was not collected using probability sampling, a representation problem occurs.

The information system provides production information, which has potential value in the decision-making process (Boehlje & Eidman, 1984). Utilization of the information system helps farmers in managing their farms and improving their performance (Hamilton & Chervany, 1981). In our case, users of “Pig Plan” are expected to have higher IT literacy and WSY (Wean per Sow a Year). Additionally, the value of the information system differs; the size of the farm is highly related to the value of the system. Therefore, manipulation of data is needed to make it representative.

The post-stratification method is employed in this study. It is a statistical method that adjusts the non-representativeness problem. It helps in resolving the non-sampling error and in getting more stable estimates (Reilly, Gelman, & Katz, 2001). This method is normally used with survey data because researchers cannot know the characteristics of the respondents before the survey data are gathered (Wang et al., 2015).

In “Pig Plan” data, the size of farms can be used as strata. When aggregating the number of weaned piglets, the average number of

	P_t	Q_t	FP_{t-r}	p_t^b	p_t^c	Y_t	S_{1t}	S_{2t}	S_{3t}
Q_t	-0.19								
	(0.010)								
FP_t	**0.66**	**0.15**							
	(0.000)	**(0.044)**							
p_t^b	**0.47**	**0.44**	**0.54**						
	(0.000)	**(0.000)**	**(0.000)**						
p_t^c	**0.57**	0.04	**0.77**	**0.37**					
	(0.000)	(0.615)	**(0.000)**	**(0.000)**					
Y_t	**0.61**	**0.42**	**0.82**	**0.80**	**0.58**				
	(0.000)	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**				
S_{1t}	0.05	0.01	-0.01	-0.09	0.1	-0.05			
	(0.504)	(0.858)	(0.846)	(0.246)	(0.182)	(0.527)			
S_{2t}	**0.2**	**-0.33**	0.02	-0.05	0.06	-0.02	**-0.33**		
	(0.006)	**(0.000)**	(0.780)	(0.498)	(0.448)	(0.759)	**(0.000)**		
S_{3t}	-0.14	**0.22**	0.02	0.09	-0.12	0.01	**-0.33**	**-0.32**	
	(0.064)	**(0.003)**	(0.830)	(0.216)	(0.101)	(0.879)	**(0.000)**	**(0.000)**	
W_{t-5}	**0.57**	**0.5**	**0.64**	**0.73**	**0.43**	**0.87**	**0.24**	**-0.16**	-0.13
	(0.000)	**(0.00)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.001)**	**(0.033)**	(0.090)

- The numbers in the parentheses are p-values

Table 8 Correlation Table of Monthly Data

Lags	P_t	Q_t	FP_t	Y_t	p_t^b	p_t^c	W_t
1	0.828	0.462	0.963	0.811	0.943	0.533	0.642
2	0.722	0.393	0.924	0.622	0.887	0.263	0.43
3	0.606	0.297	0.881	0.434	0.839	0.197	0.146
4	0.451	0.054	0.843	0.507	0.782	0.139	0.119
5	0.384	0.008	0.807	0.58	0.729	0.095	0.121

Table 9 Autocorrelation Table of Monthly Data

weaned piglets in each stratum is multiplied by the corresponding post-stratification weight as given below.

$$\bar{W}_{ps} = \sum_{h} \frac{N_h \overline{w_h}}{N} \tag{5.1}$$

where N is the population number of total farmers, N_h is the population number of farmers in strata h, and $\overline{w_h}$ is the sample mean of the number of weaned piglets in strata h.

	P_t	Q_t	FP_{t-r}	p_t^b	p_t^c	Y_t	S_{1t}	S_{2t}	S_{3t}	W_{t-21}
Q_t	**-0.37**									
	(0.000)									
FP_{t-r}	**-0.34**	**-0.23**								
	(0.000)	**(0.000)**								
p_t^b	0.06	**0.37**	**-0.79**							
	(0.264)	**(0.000)**	**(0.000)**							
p_t^c	-0.04	**-0.17**	**0.41**	**-0.45**						
	(0.392)	**(0.001)**	**(0.000)**	**(0.000)**						
Y_t	-0.05	**0.47**	**-0.86**	**0.8**	**-0.36**					
	(0.369)	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**					
S_{1t}	**-0.15**	0.04	-0.05	-0.02	**0.28**	**0.00**				
	(0.003)	(0.492)	(0.350)	(0.635)	**(0.000)**	**(1.000)**				
S_{2t}	**0.37**	**-0.23**	-0.02	**-0.12**	-0.03	0.00	**-0.34**			
	(0.000)	**(0.000)**	(0.758)	**(0.025)**	(0.582)	(1.000)	**(0.000)**			
S_{3t}	-0.06	-0.05	0.05	0.11	-0.04	0.00	**-0.34**	**-0.32**		
	(0.281)	(0.382)	(0.346)	(0.031)	(0.407)	(1.000)	**(0.000)**	**(0.000)**		
W_{t-21}	**-0.15**	**0.31**	**-0.52**	**0.44**	**-0.33**	**0.80**	**0.15**	**-0.14**	-0.08	
	(0.003)	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.004)**	**(0.005)**	(0.103)	
TO	**0.92**	**-0.3**	**-0.59**	**0.28**	**-0.13**	**0.17**	**-0.13**	**0.32**	-0.05	-0.02
	(0.000)	**(0.000)**	**(0.000)**	**(0.000)**	**(0.014)**	**(0.001)**	**(0.012)**	**(0.000)**	(0.370)	(0.731)

- The numbers in the parentheses are p-values

Table 10 Correlation Table for Weekly Data

Lags	P_t	Q_t	FP_t	p_t^b	p_t^c	W_t	TO_t
1	0.937	0.078	0.993	0.968	0.856	0.912	0.958
2	0.858	0.129	0.984	0.936	0.659	0.843	0.902
3	0.79	0.177	0.976	0.926	0.51	0.773	0.853
4	0.74	0.245	0.967	0.925	0.409	0.706	0.815
5	0.685	0.184	0.959	0.915	0.343	0.63	0.777

Table 11 Autocorrelation Table of Weekly Data

The correlation and autocorrelation tables for both monthly and weekly data are provided from Table 8 to Table 11. Most of the correlations of monthly data are similar to that of the weekly data. It should be noted that the correlation between W_{t-5} and other variables is high and significant. This implies that the number of weaned piglets is related to the future market situation. So, REH seems to be correct. In other words, farmers seemingly make decisions based on the market prediction.

It is difficult to forecast using weekly data because the number of weeks in a year varies; it is not always fifty-two weeks (Rob J Hyndman, 2014). Weeks in a year are expressed as a decimal number (00–53) using Monday as the first day of the week, typically with the first Monday of the year as day 1 of week 1 (the UK convention). When using R, the *as.Date* function was used to convert a character to date class. After that, the format function was applied, using *%W* as the conversion specification.

Year	The number of weeks	The number of dates in week 0	The number of dates in week 52	Adjustment
2011	53	2	6	Week 0 → Week 1
2012	54	1	1(53)*	Week 0 → Week 1 & Week 53 → Week 52
2013	53	6	2	Week 52 → Week 52,
2014	53	5	3	Week 52 → Week 51
2015	53	4	4	Week 0 → Week 1
2016	53	3	6	Week 0 → Week 1
2017	53	1	7	Week 0 → Week 1
2018	53	0	-	-

** In 2012, there is one day in week 53, so it is also added to week 52*
- The arrow sign (→) means the left is absorbed to the right

Table 12 Weekly Data Adjustment

Week 0 is combined with week as default. If the number of days in week 0 is greater than that of week 52, then week 51 and week 52 are merged and week 0 become week 1, as shown in Table 12.

Finally, the data are divided into three different data sets: a training set, validation set, and test set. Different from time-series analysis, machine-learning techniques need a separate validation data set to determine the meta-parameters of the model.

For monthly data, the training set is composed of data from January 2012 to February 2016 (50 months), and the validation set contains data from March 2016 to May 2017 (15 months). The test set consists of data from June 2017 to March 2018 (10 months). For weekly data, the training set is composed of data from the first week of 2012 to the eighth week of 2016 (216 weeks), and the validation set is composed of data from the

Data Type	Total Sample	Training Set	Validation Set	Test Set
Monthly data	January 2012 to March 2018 (75 months)	January 2012 to February 2016 (50 months)	March 2016 to May 2017 (15 months)	June 2017 to March 2018 (10 months)
Weekly data	week 1 2012 to week 12 2018 (324 weeks)	week 1 2012 to week 8 2016 (216 weeks)	week 9 2016 to week 20 of 2017 (64 weeks)	week 21 2017 to week 12 2018 (44 weeks)

Table 13 Sample Decomposition

ninth week of 2016 to the twentieth week of 2017 (64 weeks). The test set includes data from the twenty-first week of 2017 to the twelfth week of 2018 (44 weeks). These are summarized in Table 13.

4.2. Testing a Model based on REH

First, the models suggested in sections 4.1 and 4.2 are analyzed by econometric methods and evaluated in terms of their explanatory power. Multiple regression is used, and the assumptions of the regression are satisfied (Appendix A). Table 14 shows the results of the supply response model based on the REH. It reveals that the expected feed price affects the hog supply. When farmers expect the feed price to rise, they try to sell stocks sooner to relieve the burden of cost. The result of seasonality is consistent with the knowledge that conception rate differs across seasons (already explained in chapter 2). Supply increases in fall, compared to winter. By contrast, supply decreases in summer, compared

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	**1,855,000**	**899,200**	**2.063**	**0.043**	*
FP_t	**-1892**	**736.7**	**-2.568**	**0.013**	*
p_t^{be}	-19.67	13.33	-1.476	0.145	
p_t^{ce}	32.44	55.9	0.580	0.564	
Y_t^e	0.234	0.151	1.546	0.127	
S_{1t}	-5,347	37,160	-0.144	0.886	
S_{2t}	**-106,800**	**36,620**	**-2.917**	**0.005**	**
S_{3t}	**77,580**	**36,990**	**2.097**	**0.040**	*

*- The asterisk indicates the significance level: "***" ≡ p < 0.001; "**"≡ p < 0.01, and "*" ≡ p < 0.05*
- R-squared = 0.4834 and adjusted R-squared = 0.4294
- F-statistic is 8.955 at 7 and 67 degree of freedom, p-value = 9.599e-08

Table 14 Result of Model under REH

to winter. The value of *R-squared* is 48.34% and the value of adjusted *R-squared* is 42.94%.

From the result of the model with the aggregated farm-level data (Table 15), the influence of the number of weaned piglets at period t-5 is significant. The sign of the coefficient is as expected. Also, the seasonality effect of summer is significant. The value of *R-squared* is 49.39% and the value of adjusted *R-squared* is 46.50%. Even though this model has only one variable (except for the seasonal dummy), its explanatory power is higher. This proves that the new data are powerful.

In the combination model (Table 16), feed price is no longer significant. The number of weaned piglets and seasonality consistently influence the hog supply. This means that the number of weaned piglets deprive the feed price variable of its influence. The value of *R- squared*

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	**916,143**	**77,758**	**11.782**	**< 2e-16**	*******
W_{t-5}	**1,463**	**249**	**5.88**	**0.000**	*******
S_{1t}	-39,195	33,083	-1.185	0.240	
S_{2t}	**-126,739**	**33,544**	**-3.778**	**0.000**	*******
S_{3t}	61,014	33,506	1.821	0.073	

*- The asterisk indicates the significance level: "***" ≡ p < 0.001; "**" ≡ p < 0.01, and "*" ≡ p < 0.05*
- R-squared = 0.4939 and adjusted R-squared = 0.4650
-F-statistic is 17.08 at 4 and 70 degree of freedom, p-value = 8.121e-10

Table 15 Result of Model with Micro Data

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	940000	921500	1.020	0.311	
FP_t	-778.7	812.6	-0.958	0.341	
p_t^{be}	-20.01	12.73	-1.572	0.121	
p_t^{ce}	24.98	53.44	0.467	0.642	
Y_t^e	0.216	0.145	1.493	0.140	
S_{1t}	-24090	36130	-0.667	0.507	
S_{2t}	**-107100**	**34970**	**-3.064**	**0.003**	******
S_{3t}	**83790**	**35390**	**2.367**	**0.021**	*****
W_{t-5}	**1140**	**417**	**2.736**	**0.008**	******

*- The asterisk indicates the significance level: "***" ≡ p < 0.001; "**" ≡ p < 0.01, and "*" ≡ p < 0.05*
- R-squared = 0.5360 and adjusted R-squared = 0.4797
- F-statistic is 9.53 at 8 and 66 degree of freedom, p-value = 1.175e-08

Table 16 Result of Combination Model

is 53.60% and the value of adjusted *R-squared* is 47.97%. The values of *R-squared* and adjusted *R-squared* have increased, compared to the previous models.

By contrast, the supply response model, the model with the aggregated farm-level data, and time-series model (VAR) are compared for their forecasting accuracy. The details of the first two models are the

	REH model	Model with Farm-level Data	VAR
RMSPE	0.084	0.071	**0.068**
MAPE	0.070	0.062	**0.059**
RMSE	114,957	**96,597**	98,553
MAE	97,958	86,667	**84,168**
TPE	0.2	**0**	**0**

Table 17 Forecasting Error of REH Model, Model with Farm-level Data, and VAR

same as explained in sections 4.1 and 4.2. In addition to the two models, the VAR model will be also compared. The endogenous variables of VAR in this analysis are market price and supply.

As presented in Table 17, the forecasting error of the VAR model is the smallest among the three models. The model with farm-level data performed the second best and the model with REH was the worst. The Reason for the low forecasting accuracy of the REH model is that the strict restrictions obstruct the understanding of the nature of the market, resulting in a loose relationship with the data (Sims, 1986). The model with farm-level data seems to have underperformed, contrary to the expectation. This is because the data cover only 15% of the population. As data accumulate, it is highly possible that the model with farm-level data will eventually perform the best.

4.3. Market Prediction using Machine-learning Methods with Disaggregated Data

Six forecasting methods are compared under various conditions; different types of data, forecasting horizons, and evaluation criteria. Two of these are benchmark models: the Random Walk (RW) model and Seasonal Random Walk (SRW) model. The RW model basically assumes that people see the future price as the present price. The SRW model expects the future price to be the price for the same period from the previous year. Equation 5.1 and 5.2 are the mathematical representation of the RW and SRW models.

$$Y_t = Y_{t-1} \tag{5.1}$$

$$Y_t = Y_{t-12} \tag{5.2}$$

Additionally, there are two time-series models and two machine-learning models. The model-specification process for time-series analysis is based on the theoretical background. A time series must be stationary and choosing the lags of the variable should be based on the ACF and PACF correlogram. The variable is mostly adopted from an earlier analysis of the study with some modification. Also, exogenous variables must satisfy the assumptions of autocorrelation and multicollinearity (Andrews, Dean, Swain, & Cole, 2013). Additionally, the relationships between the dependent variables and independent

variables are linear. By contrast, the machine-learning method has no such assumptions. The variable selection process is unrestricted. All related lags are considered. For instance, price with all lags are considered 150 days[②] before the sale.

When forecasting price, the supply during the period of interest is out-of-sample. Because the input variables are unknown for the future, researchers have to rely on forecasts or the present value. However, in this study, the number of weaned piglets is used instead of the supply, thus solving the problem. The number of weaned piglets is available from 144 days before, and is highly related to the supply.

For each model of time series and machine learning, one model with a set of meta-parameters is chosen, as summarized in Table 18.
These methods will use two different data sets: monthly data and weekly data. As the machine learning methods relax assumptions of statistical methods, more complex data can be fitted into them. Hence, the machine-learning methods are expected to be superior to time-series data when using weekly data.

Each model will forecast one to five periods ahead. Forecasting performance can be revealed with a different future (Bloznelis, 2018). When time-series models forecast more than one period ahead, they use the forecasts to forecast the farther future. Therefore, the accuracy of the

② It takes about 150 days from weaning to sale.

Method	No. of models compared	Meta-parameters	Candidate values
Seasonal ARIMA	11*11*11*11=14,641	order of autoregressive term p	1, 2, ···, 11
		order of moving average term q	1, 2, ···, 11
		order of differencing d	1
		Seasonal order of autoregressive term P	1, 2, ···, 11
		Seasonal order of moving average term Q	1, 2, ···, 11
		Seasonal order of differencing D	1
Seasonal VAR	2*11*11*11*11 =29,282	Existence of the co-integration	T, F**
		order of autoregressive term p for two endogenous variables	1, 2, ···, 11
		Seasonal order of autoregressive for two endogenous variables	1, 2, ···, 11
SVR	21*21*10 = 4,410	ε-insensitive loss	$2^{(-10, \ldots, 10)}$
		Regularization parameter C	$2^{(-10, \ldots, 10)}$
		kernel parameter $\gamma = \frac{1}{\sigma^2}$	0.01, 0.02, …, 0.1
ANN	20*20*10*2 = 8,000	No. of nodes in the 1st hidden layer	1, 2, ···, 20
		No. of nodes in the 2nd hidden layer	1, 2, ···, 20
		Regularization parameter (weight decay)	0.01, 0.02, ···, 0.1
		Backpropagation method	rprop+, rprop-

** Parameters are chosen through statistical tests*
*** If true, the vector error correction model is used instead.*
The table is based on monthly data. If weekly data are used, the numbers can vary, but the same parameters are adjusted.

Table 18 Summary of Controlled Meta-parameters

time-series models is expected to decrease when forecasting farther, multiplying the errors.

Table 19 displays five types of forecast accuracy results, across forecast methods and horizons, with monthly data.

Measure	Period	RW	SRW	SARIMA	SVECM	SVR	ANN
RMSPE	1	0.096	0.096	0.065	**0.037**	0.079	0.082
	2	0.120	0.096	0.108	0.112	**0.077**	0.079
	3	0.149	0.096	0.118	0.137	**0.074**	0.114
	4	0.193	0.096	0.095	0.154	**0.071**	0.100
	5	0.208	0.096	0.143	0.158	**0.070**	0.078
MAPE	1	0.064	0.080	0.051	**0.031**	0.065	0.067
	2	0.083	0.080	0.079	0.072	**0.062**	0.062
	3	0.123	0.080	0.097	0.103	**0.060**	0.092
	4	0.166	0.080	0.084	0.123	**0.058**	0.088
	5	0.189	0.080	0.123	0.136	**0.057**	0.067
RMSE	1	389.17	423.20	277.50	**155.15**	368.29	353.50
	2	489.52	423.20	432.71	424.52	**353.29**	355.65
	3	624.26	423.20	480.31	529.33	**335.33**	510.96
	4	795.93	423.20	421.86	598.04	**321.43**	437.35
	5	859.95	423.20	573.88	622.46	**315.28**	354.77
MAE	1	267.60	352.80	220.93	**130.82**	293.33	292.34
	2	345.70	352.80	331.18	291.74	**279.92**	277.81
	3	523.00	352.80	405.61	417.70	**268.73**	404.68
	4	701.00	352.80	364.33	499.56	**256.48**	386.34
	5	800.90	352.80	513.25	554.19	**252.14**	301.02
TPE	1	**0.133**	0.200	0.267	0.200	0.267	0.200
	2	**0.133**	0.200	0.267	0.200	0.200	0.200
	3	0.267	0.200	0.200	**0.067**	0.200	0.267
	4	0.333	0.200	0.333	**0.133**	0.200	0.267
	5	0.267	0.200	0.200	**0.133**	0.200	0.267

- The best result under each measure and period is bold-faced

Table 19 Forecasting Errors Using Monthly Data

When forecasting the nearest horizon, the time-series analysis method dominates the machine-learning method. SVECM performs the best when forecasting a week ahead, regardless of the error measures, except for TPE. As the forecasting horizon increases, the time-series method

loses its dominance to the machine-learning method. Errors in the time-series method increase dramatically when forecasting ahead. This can be explained by the fact that time-series analysis forecasts the far future, thus duplicating errors. Meanwhile, the forecasting accuracy of machine-learning methods is stable. However, TPE does not follow most of the results of error measures. This is because monthly data show strong seasonality. Although the value of the error for time-series analysis is larger than that in the machine-learning method, the direction of the forecast is more accurate.

Table 20 displays five types of forecast error measures across forecasting methods and horizons, with weekly data. From the results of the forecast with monthly data, time-series analysis methods dominate machine-learning methods in the near horizon. With errors based on absolute value, SVAR dominates the machine-learning method, when forecasting one to three weeks ahead; with errors based on mean squared value, VAR performs the best only when forecasting a week ahead. Because mean squared value assigns greater weightage to poor forecasts, the result means that SVAR produces huge misprediction at some point, even though the error of SVAR is low. It implies that using VAR can be risky, because it occasionally results in incorrect forecasts. TPE shows consistent results with other criteria, possibly because weekly data have weaker seasonality than monthly data.

Measure	Period	RW	SRW	SARIMA	SVAR	SVR	ANN
RMSPE	1	0.045	0.103	0.080	**0.028**	0.040	0.065
	2	0.074	0.103	0.187	0.058	0.073	**0.048**
	3	0.093	0.103	0.118	0.079	0.090	**0.070**
	4	0.111	0.103	0.144	0.095	0.090	**0.068**
	5	0.126	0.103	0.159	0.102	0.090	**0.068**
MAPE	1	0.033	0.085	0.069	**0.020**	0.029	0.054
	2	0.054	0.085	0.159	0.042	0.063	**0.039**
	3	0.068	0.085	0.083	**0.058**	0.075	0.060
	4	0.083	0.085	0.104	0.069	0.075	**0.055**
	5	0.092	0.085	0.119	0.076	0.075	**0.055**
RMSE	1	207.34	488.21	411.47	**130.65**	195.16	306.50
	2	334.74	488.21	779.00	258.31	338.23	**244.21**
	3	420.81	488.21	502.12	344.50	409.58	**333.48**
	4	494.72	488.21	611.70	408.18	409.58	**321.57**
	5	554.77	488.21	675.74	440.64	409.58	**321.57**
MAE	1	151.06	397.70	335.74	**95.90**	142.16	257.70
	2	249.88	397.70	686.16	**193.23**	299.61	195.89
	3	312.90	397.70	374.65	**267.96**	351.21	287.53
	4	379.02	397.70	469.96	315.30	351.21	**268.69**
	5	417.07	397.70	535.84	345.76	351.21	**268.69**
TPE	1	0.455	0.409	0.364	**0.182**	0.235	0.412
	2	0.614	0.409	0.500	**0.318**	0.529	0.353
	3	0.568	0.409	0.364	0.409	0.471	**0.294**
	4	0.614	0.409	0.409	0.455	0.471	**0.324**
	5	0.591	0.409	0.432	0.432	0.471	**0.324**

*- The best result under each measure and period is boldfaced and marked * and the second best is boldfaced only*

Table 20 Forecasting Errors Using Weekly Data

Chapter 5. Conclusion and Discussion

Based on the market prediction background, this study tries to resolve two controversial issues. One is the issue of REH, a fundamental theory of human behavior in economics. Models based on REH have been criticized by Sims (1986) and his followers for its poor forecasting ability, resulting in nonoptimal and unreliable relationships with the data. By using the information system data, the REH model was tested in terms of its explanatory and predictive powers.

The REH model explained the market poorly, compared with the model based on farm-level data. The number of weaned piglets, used in the model based on farm-level data, directly translate to the supply after five months. This variable is more effective in explaining the supply than the REH. Also, when the variables of the two models are combined, some variables from the REH model become insignificant. This shows that micro data diluted some of the impact of other variables.

The forecast error of the REH model was larger than that of the model based on farm-level data and VAR. This proves that the theory-driven model's forecasting ability is inferior to the data-driven model due to its restriction on human behavior. Even though the weaned piglets are expected to constitute the future hog supply owing to the biological cycle,

the result of the model using farm-level data performed worse than VAR. This can be resolved as data accumulate, because in this model only 15% of the population data is used.

The other issue relates to the machine-learning method. Even though the machine-learning method has been highlighted for its utility, its performance has been questioned. In this study, forecasting methods were compared across error measures and horizons. Regardless of the data used, time-series methods are better when forecasting the near horizon, while machine-learning methods perform better when forecasting the far horizon. This result can be explained using various approaches. Strong seasonality exists in monthly data and machine-learning methods do not recognize data as time series. In addition, the time-series data uses forecasts to forecast farther. For instance, when forecasting P_{t+3}, time-series data uses forecasts of P_{t+1} to forecast P_{t+2}, and then uses forecasts of P_{t+2} to forecast P_{t+3}. Therefore, the forecast error is accumulated when forecasting farther. Because the number of the weaned piglets is adopted instead of the hog supply, it allows to forecast at most 144 days ahead. The result of the machine-learning method seems to be almost stable, regardless of the forecasting horizon, partly affected by the availability of the variable that is related to the forecast, 144 days ahead.

It is difficult to conclude that one methodology always performs the

best. Rather, time-series analysis is recommended when forecasting the near future. It confirms that time-series analysis is a useful tool for short-term forecasting. By contrast, machine-learning methods are better when forecasting the far future. It means that the relationship between the present and the far future is not linear. When forecasting a weekly horizon, the time-series method can be risky, as it forecasts with volatility.

The study has both theoretical and practical implications. This study predicts the agricultural market using farm-level data. As data are compiled, it challenges the strict assumptions of economic theory. This study shows the poor explanatory and forecasting ability of REH, and the power of application of micro-level data in solving the problem of REH. The conventional assumptions about information accessibility and usability can be eliminated with the adoption of such data, because the data reflect the actual behavior of humans.

Microeconomists have focused on the market behavior over the individual's behavior and regard the research of individual behavior as a field of business study. In the information age, microeconomics can be criticized for ignoring individual differences, as new classical macroeconomists have criticized the lack of micro-foundation of macroeconomics in the 1970s (Brunner & Meltzer, 1983).

Even though some studies have tried to forecast the market price with machine-learning methods, most of the studies used only aggregated

data. This study uses disaggregated farm-level data. It allows for the reflection of the different decisions of the farmers, and the machine-learning methods fully utilize the complexity of the data.

Empirically, it gives some hints for managers and policymakers about what method is to be used to predict the market. Precise price forecasting can increase market stability. It can benefit both farmers and consumers by reducing risks of decision-making and by securing market price.

The quality of data was emphasized along with the quantity of data. The reason for using data after 2012 is that the circumstances for the hog industry changed in 2012. In 2011, the foot-and-mouth disease (FMD) proliferated and had a profound effect on hog supply. A different trend was captured after the outbreak of FMD. Especially, before 2012, injection of the vaccination was not mandatory; if FMD occurred, all pigs in the same region had to be slaughtered. After 2012, vaccination became compulsory; If FMD breaks out, only the infected livestock is slaughtered. Besides, before 2012, there were many missing values in the "Pig Plan" data and its users were limited.

This study has certain limitations. It deals with the hog price. However, the characteristics of agricultural products vary widely, unlike the case with manufactured products. Consequently, this result might not be applicable to the other crops. Therefore, research with different data

is needed to generalize the forecasting result.

Moreover, newer forecasting methodologies were not included in this study. Only basic models are considered, because applying the latest techniques to the hog sector was not the main goal of this study. Many attempts have been made to combine machine-learning models and time-series models, similar to the older trends of combining economic models with time-series models. Later, these models can be employed to obtain more rigorous results. Also, recursive neural network can be more appropriate than time-series data. This work is left for future studies.

Using micro data (private data) seems promising, in terms of research and stabilization of the market. However, there are many obstacles for the actual implementation, such as data privacy issues and unwillingness of firms to share data. It will be helpful to analyze the social benefits of using such data.

Bibliography

Adya, M., & Collopy, F. (1998). How effective are neural networks at forecasting and prediction? A review and evaluation. *Journal of Forecasting, 17*(5-6), 481-495.

Andrews, B. H., Dean, M. D., Swain, R., & Cole, C. (2013). Building ARIMA and ARIMAX models for predicting long-term disability benefit application rates in the public/private sectors. *Society of Actuaries*, 1-54.

Antonovitz, F., & Green, R. (1990). Alternative Estimates of Fed Beef Supply Response to Risk. *American Journal of Agricultural Economics, 72*(2), 475-487. doi:10.2307/1242351

Arzac, E. R., & Wilkinson, M. (1979). A quarterly econometric model of United States livestock and feed grain markets and some of its policy implications. *American Journal of Agricultural Economics, 61*(2), 297-308.

Awad, M., & Khanna, R. (2015). *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Berkeley, CA: Apress.

Bloznelis, D. (2018). Short-term salmon price forecasting. *Journal of Forecasting, 37*(2), 151-169. doi:10.1002/for.2482

Boehlje, M. D., & Eidman, V. R. (1984). *Farm management*: J. Wiley.

Bragoli, D. (2017). Now-casting the Japanese economy. *International Journal of Forecasting, 33*(2), 390-402. doi:https://doi.org/10.1016/j.ijforecast.2016.11.004

Brandt, J. A., & Bessler, D. A. (1981). Composite forecasting: An application with US hog prices. *American Journal of Agricultural Economics, 63*(1), 135-140.

Bruhns, A., Deurveilher, G., & Roy, J.-S. (2005). *A non linear regression model for mid-term load forecasting and improvements in seasonality.* Paper presented at the Proceedings of the 15th power systems computation conference.

Brunner, K., & Meltzer, A. (1983). *Econometric policy evaluation. A critique*. Paper presented at the Theory, Policy, Institutions: Papers from the

Carnegie-Rochester Conferences on Public Policy.

Chakraborty, K., Mehrotra, K., Mohan, C. K., & Ranka, S. (1992). Forecasting the behavior of multivariate time series using neural networks. *Neural networks, 5*(6), 961-970.

Choe, Y. C., & Lee, M. S. (2010). Information technology payoff: A panel data application to swine farm in Korea. *Americas Conference on Information Systems*.

Choi, J. S. (2016). Evaluation of Estimation and Forecast Accuracy on Retail Meat Prices by Seasonal Time-Series Models. [Evaluation of Estimation and Forecast Accuracy on Retail Meat Prices by Seasonal Time-Series Models]. *Korean Journal of Food Marketing Economics, 33*(1), 1-31.

Chow, G. C. (1989). Rational Versus Adaptive Expectations in Present Value Models. *The Review of Economics and Statistics, 71*(3), 376-384. doi:10.2307/1926893

Danziger, L. (1987). Inflation, fixed cost of price adjustment, and measurement of relative-price variability: Theory and evidence. *The American Economic Review, 77*(4), 704-713.

Figlewski, S., & Wachtel, P. (1981). The formation of inflationary expectations. *The Review of Economics and Statistics*, 1-10.

Fornaro, P. (2016). Predicting Finnish economic activity using firm-level data. *International Journal of Forecasting, 32*(1), 10-19. doi:https://doi.org/10.1016/j.ijforecast.2015.04.002

Fulford, A. J., Rayco-Solon, P., & Prentice, A. M. (2006). Statistical modelling of the seasonality of preterm delivery and intrauterine growth restriction in rural Gambia. *Paediatric and perinatal epidemiology, 20*(3), 251-259.

Goodwin, T. H., & Sheffrin, S. M. (1982). Testing the Rational Expectations Hypothesis in an Agricultural Market. *The Review of Economics and Statistics, 64*(4), 658-667. doi:10.2307/1923950

Guidolin, M., & Thornton, D. L. (2018). Predictions of short-term rates and the expectations hypothesis. *International Journal of Forecasting, 34*(4), 636-664. doi:https://doi.org/10.1016/j.ijforecast.2018.03.006

Haltiwanger, J., & Waldman, M. (1985). Rational expectations and the limits

of rationality: An analysis of heterogeneity. *The American Economic Review, 75*(3), 326-340.

Hamilton, S., & Chervany, N. L. (1981). Evaluating information system effectiveness-Part I: Comparing evaluation approaches. *MIS Quarterly*, 55-69.

Hamm, L., & Brorsen, B. W. (1997). Forecasting Hog Prices with a Neural Network. *Journal of Agribusiness, 15*(1), 37-54. doi:http://purl.umn.edu/90646

Harchaoui, T. M., & Janssen, R. V. (2018). How can big data enhance the timeliness of official statistics?: The case of the U.S. consumer price index. *International Journal of Forecasting, 34*(2), 225-234. doi:https://doi.org/10.1016/j.ijforecast.2017.12.002

Hendry, D. F., & Muellbauer, J. N. (2018). The future of macroeconomics: macro theory and models at the Bank of England. *Oxford Review of Economic Policy, 34*(1-2), 287-328.

Holt, M. T., & Johnson, S. R. (1988). Supply Dynamics in the U.S. Hog Industry. *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie, 36*(2), 313-335. doi:10.1111/j.1744-7976.1988.tb03278.x

Hommes, C. (2011). The heterogeneous expectations hypothesis: Some evidence from the lab. *Journal of Economic Dynamics and Control, 35*(1), 1-24.

Hyndman, R. J. (2010). Forecasting with long seasonal periods. Retrieved from https://robjhyndman.com/hyndsight/longseasonality/

Hyndman, R. J. (2014). Forecasting weekly data. Retrieved from https://robjhyndman.com/hyndsight/forecasting-weekly-data/

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*: OTexts.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679-688. doi:https://doi.org/10.1016/j.ijforecast.2006.03.001

Iida, R., & Koketsu, Y. (2013). Interactions between climatic and production factors on returns of female pigs to service during summer in Japanese commercial breeding herds. *Theriogenology, 80*(5), 487-493.

Jung, M., & Kim, H. (2011). 쇠고기・돼지고기 수급구조 분석 및 정책 시뮬레이션 모형 개발. *한국농촌경제연구원 기본연구보고서*, 1-251.

Kaastra, I., & Boyd, M. (1996). Designing a neural network for forecasting financial and economic time series. *Neurocomputing, 10*(3), 215-236.

Kim, S. (1998). fluctuation of farm price of hog and the cause. *Journal of Rural Development, 21*(1), 19-31.

Klein, P. (2000). Using the generalized Schur form to solve a multivariate linear rational expectations model. *Journal of Economic Dynamics and Control, 24*(10), 1405-1423. doi:https://doi.org/10.1016/S0165-1889(99)00045-7

Lessmann, S., & Voß, S. (2017). Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy. *International Journal of Forecasting, 33*(4), 864-877. doi:https://doi.org/10.1016/j.ijforecast.2017.04.003

Leuthold, R., MacCormick, A., Schmitz, A., & Watts, D. (1970). Forecasting daily hog prices and quantities: A study of alternative forecasting techniques. *Journal of the American Statistical Association, 65*(329), 90-107.

Lovell, M. C. (1986). Tests of the rational expectations hypothesis. *The American Economic Review, 76*(1), 110-124.

Lucas, R. (1976). Econometric policy evaluation: A critique. The Phillips curve and labour markets. K. Brunner and AH Meltzer. *Carnegie-Rochester Conference Series on Public Policy, 1*(1), 19-46.

Marcjasz, G., Uniejewski, B., & Weron, R. (2018). On the importance of the long-term seasonal component in day-ahead electricity price forecasting with NARX neural networks. *International Journal of Forecasting*. doi:https://doi.org/10.1016/j.ijforecast.2017.11.009

Meilke, K. D., Zwart, A. C., & Martin, L. J. (1974). NORTH AMERICAN HOG SUPPLY: A COMPARISON OF GEOMETRIC AND POLYNOMIAL DISTRIBUTED LAG MODELS*. *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie, 22*(2), 15-30. doi:10.1111/j.1744-7976.1974.tb00925.x

Muth, J. F. (1961). Rational expectations and the theory of price movements.

Econometrica: Journal of the Econometric Society, 315-335.

Nerlove, M. (1958). Distributed lags and estimation of long-run supply and demand elasticities: Theoretical considerations. *Journal of Farm Economics, 40*(2), 301-311.

Nikolopoulos, K., Goodwin, P., Patelis, A., & Assimakopoulos, V. (2007). Forecasting with cue information: A comparison of multiple regression with alternative forecasting approaches. *European Journal of Operational Research, 180*(1), 354-368.

Parsley, D. C. (1996). Inflation and relative price variability in the short and long run: new evidence from the United States. *Journal of Money, Credit and Banking, 28*(3), 323-341.

Peel, D., & Meyer, S. (2002). Cattle price seasonality. *Managing for Today's Cattle Market and beyond*.

Powell, B., Nason, G., Elliott, D., Mayhew, M., Davies, J., & Winton, J. (2017). Tracking and modelling prices using web-scraped price microdata: towards automated daily consumer price index forecasting. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 0*(0). doi:doi:10.1111/rssa.12314

Prescott, D. M., & Stengos, T. (1987). Bootstrapping Confidence Intervals: An Application to Forecasting the Supply of Pork. *American Journal of Agricultural Economics, 69*(2), 266-273. doi:10.2307/1242276

Rayco-Solon, P., Fulford, A. J., & Prentice, A. M. (2005). Differential effects of seasonality on preterm birth and intrauterine growth restriction in rural Africans–. *The American journal of clinical nutrition, 81*(1), 134-139.

Reilly, C., Gelman, A., & Katz, J. (2001). Poststratification without population level information on the poststratifying variable with application to political polling. *Journal of the American Statistical Association, 96*(453), 1-11.

Reutlinger, S. (1966). Short-Run Beef Supply Response. *American Journal of Agricultural Economics, 48*(4_Part_I), 909-919. doi:10.2307/1236621

Runkle, D. E. (1991). Are Farrowing Intentions Rational Forecasts? *American Journal of Agricultural Economics, 73*(3), 594-600. doi:10.2307/1242812

Shahwan, T., & Odening, M. (2007). Forecasting agricultural commodity prices using hybrid neural networks. In *Computational intelligence in economics and finance* (pp. 63-74): Springer.

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289-310.

Sims, C. A. (1986). Are forecasting models usable for policy analysis? *Quarterly Review*(Win), 2-16.

Taylor, J. W., McSharry, P. E., & Buizza, R. (2009). Wind power density forecasting using ensemble predictions and time series models. *IEEE Transactions on Energy Conversion, 24*(3), 775.

Tryfos, P. (1974). Canadian supply functions for livestock and meat. *American Journal of Agricultural Economics, 56*(1), 107-113.

Wallis, K. F. (1980). Econometric Implications of the Rational Expectations Hypothesis. *Econometrica, 48*(1), 49-73. doi:10.2307/1912018

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting, 31*(3), 980-991. doi:https://doi.org/10.1016/j.ijforecast.2014.06.001

Weizsäcker, G. (2010). Do we follow others when we should? A simple test of rational expectations. *American Economic Review, 100*(5), 2340-2360.

Witt, C. A., & Witt, S. F. (1989). Measures of forecasting accuracy — turning point error v size of error. *Tourism management, 10*(3), 255-260. doi:https://doi.org/10.1016/0261-5177(89)90087-3

Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M.-J. (2017). Big data in smart farming–a review. *Agricultural Systems, 153*, 69-80.

Ye, M., Zyren, J., & Shore, J. (2005). A monthly crude oil spot price forecasting model using relative inventories. *International Journal of Forecasting, 21*(3), 491-501.

Appendix

Appendix A. Test of Assumptions

Appendix A.1. Test Result of Model under REH

Studentized Breusch-Pagan Test		Durbin-Watson Test	
BP statistic	13.65	DW statistic	1.90
p-value	0.058	p-value	0.169

Value of Variance Inflation Factor						
FP_{t-r}	p_t^{be}	p_t^{ce}	Y_t^e	S_{1t}	S_{2t}	S_{3t}
5.283	4.203	1.529	3.058	1.734	1.625	1.657

Appendix A.2. Test Result of Model with Micro Data

Studentized Breusch-Pagan Test		Durbin-Watson Test	
BP statistic	7.43	DW statistic	1.69
p-value	0.115	p-value	0.050

Value of Variance Inflation Factor			
S_{1t}	S_{2t}	S_{3t}	W_{t-5}
1.466	1.454	1.450	1.033

Appendix A.3. Test Result of Combination Model

Studentized Breusch-Pagan Test		Durbin-Watson Test	
BP statistic	11.57	DW statistic	1.77
p-value	0.172	p-value	0.067

Value of Variance Inflation Factor							
FP_{t-r}	p_t^{be}	p_t^{ce}	Y_t^e	S_{1t}	S_{2t}	S_{3t}	W_{t-5}
7.051	4.203	1.532	3.065	1.799	1.625	1.664	2.980

요약 (국문초록)

정보시스템 데이터를 활용한 시장 예측에 관한 연구

Data-driven Agricultural Market Prediction
- Using Data from Hog-farm Management Information System -

기술의 발전과 더불어 개개인의 의사결정이 데이터로 축적되면서 이러한 데이터를 활용하여 과거에는 풀지 못하였던 다양한 문제들을 검증할 수 있게 되었다. 본 연구는 시장예측과 관련된 선행 연구에 있어서 두 개의 주요한 연구 질문에 대한 답을 구하고자 한다. 첫 번째 주요한 문제는 합리적 기대가설이 이론적으로는 널리 활용되어 왔지만 너무 강한 가정을 가지고 있기 때문에 실제 경제 예측을 잘못한다는 점이다. 본 연구에서 합리적 기대 가설을 활용한 모델과 개개인의 의사결정을 합한 데이터를 활용한 모델을 비교해보니 정보 시스템 데이터를 활용한 모델이 더 뛰어난 설명력과 예측력을 가지는 것을 확인 할 수 있었다. 두 번째로 주요한 질문은 개개인의 의사결정을 담는 데이터를 활용하여 시장 예측을 할 때 가장 유용한 방법론이 어떤 것인가에 대한 질문이다. 과거에 시장 가격과 시장 공급량과 같이 합해진 데이터를 사용할 때는 시계열 방법론이 예측을 잘 한다는 결과가 있었다. 그러나 시장 단위가 아닌 개인 단위의 데이터는 더욱 복잡하고

비선형적인 관계가 나타날 수 있기 때문에 본 연구에서는 머신러닝 방법을 활용하여 시계열 방법과 비교해 보았다. 그 결과 단기적으로는 시계열 방법이 장기적으로는 머신러닝 방법이 우세하다는 결과를 얻을 수 있었다.

주요어 : 합리적 기대 가설, 가격 예측, 인공신경망,
서포트벡터회귀, SARIMA, VAR
학 번 : 2016-21492