Vegetable Price Prediction Using Atypical Web-Search Data

Do-il Yoo
Department of Agricultural Economics
Chungbuk National University
Email: d1yoo@chungbuk.ac.kr

Selected Paper prepared for presentation at the 2016 Agricultural & Applied Economics Association Annual Meeting, Boston, Massachusetts, July 31-August 2

1. Introduction

In vegetable market, reliable price prediction is expected to prevent loss of social welfare caused by excess supply or excess demand. For example, by referring to predicted future price, farmers may produce less vegetable beforehand in the excess supply market, where price is expected to drop. And, farmers' efficient quantity adjustment can save potential social costs of unshipped product waste landfills, long-term storage, farm subsidies, and etc. Thus, it's necessary to predict vegetable price as accurate as possible.

Traditionally, a considerable previous literature relies on time series or neural network models in predicting price in the sense that past prices may impact on current and future prices. Thanks to innovative information technology, recent Big-Data boom receives huge attention as it is possible to analyze large dataset gathered from online websites like Google and social network services (SNS) such as blogs, Twitter, Facebook, and etc. Associated literature pays attention to the impact of atypical web-search data composed of specific lexicon on relevant product sales or prices assuming that those lexicons reflect consumers' psychology in making economic decisions. Representatively, Google search engine query data are used to predict economic indicators such as automobile sales, unemployment claims, consumer sentiment, and gun sales (Choi and Varian, 2012; Scott and Varian, 2013). Bollen et al. (2011) predicts the stock market by analyzing the influence of public Twitter mood on the value of the Dow Jones Industrial Average.

Though Big-Data issue is actively spreading in the field of finance, marketing, and economics, studies concerning agricultural economics are relatively rare. Rare studies result from the fact that agricultural products market is more uncertain and unpredictable than other industrial products market; agricultural products are easily perishable and frequently affected by climate factors, leading to fluctuating prices. Therefore, it would be timely to introduce atypical web-search data analysis into the field of agricultural economics. We pay attention to

the impact of lexicons concerning vegetables on websites on vegetable prices.

The object of this study is to develop vegetable price prediction model with higher prediction power. Based on the typical time-series models, we pay attention to the role of atypical web-search data obtained from on-line websites. Here, we believe that such atypical data could provide more robust price prediction.

To do so, we depend on the Bayesian structural time series (BSTS) model suggested by Scott and Varian (2013). While typical time-series models focus on the relations between current prices and lagged prices, structural time series models could be more useful in the sense that explanatory variables impacting prices are introduced in the structural form (Harvey and Shephard, 1993). In addition, the Bayesian approach is widely used to provide better prediction concerning random walk by using updated posterior information from prior information of random walk (Koop, 2003).

The paper is organized as follows. Section 2 presents both conceptual and empirical BSTS models for vegetable price prediction. Section 3 presents an application of the approach to three vegetables of dried red pepper, garlic, and onion in Korean wholesale market. Predicted price results are reported in section 4. Finally, section 5 concludes.

2. Bayesian Structural Time Series (BSTS) Model

Based on the state space form, where unobserved latent variables are considered as state variables, a typical conceptual model using BSTS is composed of two equations as follows:

$$y_t = Z_t^T \alpha_t + \varepsilon_t, \qquad \text{where} \quad \varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right) \tag{1}$$

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t \qquad , \text{where} \quad \eta_t \sim N\left(0, Q_t\right) \tag{2}$$

Equation (1) is called as an observation equation, linking observable time-series data y_t with unobserved latent variables (state variables) α_t. And, equation (2) is called as a transition equation describing the law of motion between the current state variables α_t and the next state variables α_{t+1}. In (1), Z_t is a vector including explanatory variables and parameters. In (2), T_t corresponds to a transition matrix accounting for relation between α_t and α_{t+1}, and R_t is a vector including parameters. Both ε_t and η_t are random noises following the Gaussian distribution with zero mean and the variance σ_ε^2 and Q_t, respectively (Harvey and Peters, 1990; Scott and Varian, 2013a).

For each vegetable $i = \{garlic, onion, drp\}$[1], equations (1) and (2) can be specified with concepts of trend μ_t^i and seasonality τ_t^i for price time-series y_t^i as follows:

$$y_t^i = \mu_t^i + \tau_t^i + \beta^{iT} x_t^i + \varepsilon_t^i \tag{3}$$

$$\mu_t^i = \mu_{t-1}^i + \delta_{t-1}^i + u_t^i \tag{4}$$

$$\delta_t^i = \delta_{t-1}^i + v_t^i \tag{5}$$

$$\tau_t^i = -\sum_{s=1}^{S^i-1} \tau_{t-s}^i + w_t^i \tag{6}$$

, where x_t^i is a vector including explanatory variables impacting y_t^i with its associated parameter vector β_t^i. In equation (4) and (5), δ_t^i is the slope of trend μ_t^i. In (6), S^i indicates the number of seasons considered in the model for i-vegetable. Except for equation (3), the other equations (4) ~ (6) account for typical time-series models. Through equation (3)

[1] drp indicates dried red pepper.

~ (6), $\left(\varepsilon_t^i, u_t^i, v_t^i, w_t^i\right)$ are also assumed to be Gaussian random noises with time-invariant variances $\left(\sigma_{\varepsilon^i}^2, \delta_{u^i}^2, \delta_{v^i}^2, \delta_{w^i}^2\right)$, respectively.

Now, it's necessary to distinct equation (3) by types of explanatory variables. That is to say, for our empirical analysis, we need evaluate which approaches can provide better price prediction with and without atypical web-search data concerning i-vegetable. Under $\left\{C_t^i, A_t^i\right\} \in x_t^i$, let's consider C_t^i is a vector composed of climate factors for i-vegetable. Also, let's consider A_t^i be a vector including atypical indexes obtained from atypical web-search data for i-vegetable. Holding equations (4) ~ (6) same, equation (3) is specified into three empirical models by the type of x_t^i as follows:

$$y_t^i = \mu_t^i + \tau_t^i + \varepsilon_t^i \tag{7}$$

$$y_t^i = \mu_t^i + \tau_t^i + \xi^{iT} C_t^i + \varepsilon_t^i \tag{8}$$

$$y_t^i = \mu_t^i + \tau_t^i + \xi^{iT} C_t^i + \psi^{iT} A_t^i + \varepsilon_t^i \tag{9}$$

, where ξ^i and ψ^i are parameter vectors associated with climate factors and atypical web-search data, respectively for i-vegetable. Also, equations (7) ~ (9) are named as 'BSTS-I', 'BSTS-II', and 'BSTS-III'. Then, BSTS-I is a benchmark model for comparing other models BSTS-II and BSTS-III. As seen in equation (7), BSTS-I has no explanatory variables in its form, implying a pure time-series model considering only trend and seasonality. BSTS-II is a BSTS model with only climate factors, whose impacts are assumed to impact vegetable price volatility through unstable demand and supply due to climate volatility. Finally, BSTS-III is a BSTS model considering both climate factors and atypical indexes using atypical web-search data. Further details concerning C_t^i and A_t^i are presented in section 3.

Estimation method for BSTS models depend on stochastic estimation. Through equations (7) ~ (9), parameters associated with models are $\beta^i = \{\xi^i, \psi^i\}$ and $\sigma^2_{\varepsilon^i}$ for each vegetable i. Their prior probability distributions $p(\beta^i)$ and $p(\sigma^2_{\varepsilon^i})$ are assumed to follow the Gaussian and the inverse Gamma distributions, respectively as follows (Koop, 2003; Scott and Varian, 2013a):

$$\beta^i \mid \sigma^2_{\varepsilon^i} \sim N(o, \sigma^2_{\varepsilon^i}\Omega^i) \tag{10}$$

$$\frac{1}{\sigma^2_{\varepsilon^i}} \sim G\left(\frac{v^i}{2}, \frac{ss^i}{2}\right) \tag{11}$$

, where Ω^i is a prior information matrix with $\Omega^{i-1} = \left(X^{iT}X^i + diag\left(X^{iT}X^i\right)\right)/2n^i$ and $X^{iT} = \left[x^i_1, \cdots, x^i_{n^i}\right]$ when $x^i_t = \{C^i_t, A^i_t\}$. Also, v^i and ss^i indicate a prior sample size and the prior sum of squares for i-vegetable (Scott and Varian, 2013b).[2]

Due to the properties of conjugacy in the Gaussian and the inverse Gamma distribution, the posterior probability distributions for parameters $p(\beta^i \mid \sigma^2_{\varepsilon^i}, y^i_1, \ldots, y^i_{n^i})$ and $p(\sigma^2_{\varepsilon^i} \mid y^i_1, \ldots, y^i_{n^i})$ also follow same distributions with prior distributions as follows (DeGroot, 2004):

$$\beta^i \mid \sigma^2_{\varepsilon^i}, y^i_1, \ldots, y^i_{n^i} \sim N\left(\left(X^{iT}X^i + \left(\Omega^i\right)^{-1}\right)^{-1}\left(X^{iT} \cdot \left[y^i_1, \ldots, y^i_{n^i}\right]^T\right), \sigma^2_{\varepsilon^i}\left(X^{iT}X^i\right) + \left(\Omega^i\right)^{-1}\right) \tag{12}$$

[2] Further details are encouraged to refer to Scott and Varian (2013b).

$$\frac{1}{\sigma_{\varepsilon^i}^2} \mid y_1^i,\ldots,y_{n^i}^i \sim G\left(\begin{array}{l}\dfrac{v^i+n^i}{2},\\ \dfrac{\left[\begin{array}{l} ss^i+\left[y_1^i,\ldots,y_{n^i}^i\right]\left[y_1^i,\ldots,y_{n^i}^i\right]^T \\ -\left(\left(X^{iT}X^i+\left(\Omega^i\right)^{-1}\right)^{-1}\left(X^{iT}\cdot\left[y_1^i,\ldots,y_{n^i}^i\right]^T\right)\right)^T \\ \cdot\left(X^{iT}X^i+\left(\Omega^i\right)^{-1}\right) \\ \cdot\left(X^{iT}X^i+\left(\Omega^i\right)^{-1}\right)^{-1}\left(X^{iT}\cdot\left[y_1^i,\ldots,y_{n^i}^i\right]^T\right)\end{array}\right]}{2}\end{array}\right) \tag{13}$$

Following Durbin and Koopman (2002), the posterior probability distributions $p\left(\beta^i \mid \sigma_{\varepsilon^i}^2, y_1^i,\ldots,y_{n^i}^i\right)$ and $p\left(\sigma_{\varepsilon^i}^2 \mid y_1^i,\ldots,y_{n^i}^i\right)$ are estimated by the Markov chain Monte Carlo (MCMC) simulation using Gibbs sampling.

Denoting $\tilde{y}^i$ as a price prediction and $\Phi^i=\left\{\beta^i,\left(\mu^i,\delta^i,\tau^i\right),\left(\sigma_{\varepsilon^i}^2,\sigma_{u^i}^2,\sigma_{v^i}^2,\sigma_{w^i}^2\right),\alpha^i\right\}$ as a combined parameter vector across all equations for i-vegetable, the posterior predictive distribution is derived from the following equation:

$$p\left(\tilde{y}^i \mid y_1^i,\ldots,y_{n^i}^i\right)=\int p\left(\tilde{y}^i \mid \Phi^i\right)p\left(\Phi^i \mid y_1^i,\ldots,y_{n^i}^i\right)d\Phi^i \tag{14}$$

, implying Bayes' theorem. Empirically, equation (14) is obtained by calculating $E\left[\tilde{y}^i \mid y_1^i,\ldots,y_{n^i}^i\right]$ based on randomly derived Φ^i using Monte Carlo estimation.

3. Data

Conceptual and empirical models developed in section 2 are applied to the Korean wholesale vegetable markets for garlic, onion, and dried red pepper at the monthly level. Reminding our

goal is to provide better vegetable price prediction across BSTS models, we specify associated explanatory variables for each vegetable i.

First, climate factors in C_t^i includes temperature $temp_t^i$, minimum temperature $\min temp_t^i$, precipitation $precip_t^i$, sunshine amount sun_t^i, and their square terms.

$$C_t^i = \left[temp_t^i, temp_t^{i2}, \min temp_t^i, \min temp_t^{i2}, precip_t^i, precip_t^{i2}, sun_t^i, sun_t^{i2} \right] \quad (15)$$

Here, square-terms are used for reflecting climate volatility instead of each climate factor's variance terms, leading to non-linear models. All values are averaged values by month as we predict monthly vegetable prices. Descriptive statistics for climate factors, prices, and quantities for each vegetable are described from <Table 1> to <Table 3>. Note that all averaged values for each climate factor for i-vegetable are calculated from chief producing districts for each vegetable as illustrated in <Figure 1> ~ <Figure 3>.

<Figure 1 ~ Figure 3, here>

<Table 1 ~ Table 3, here>

Second, atypical indexes in A_t^i are derived from atypical web-search data obtained from various on-line websites including SNS. We suggest five atypical indexes according to recent text-mining approaches widely used in the Big-Data research, reflecting consumers' attention on three vegetables from SNS websites and major portal sites such as Google and Naver in Korea. Specifically, using text mining program 'Textom' and 'UNICET 6', we gather associate web-search keywords. Then, we make simple query data measuring

frequency on websites and Term Frequency – Inverse Document Frequency (TF-IDF) considering weights of core keywords on websites (Salton and McGill, 1983). So, five atypical indexes are as follows:

$$A_t^i = \left[\text{info}_t^i, search_t^i, unb_t^i, pec_t^i, link_t^i \right] \tag{16}$$

, where info_t^i is an index for information extracted from web documents using text-mining approach, implying a total amount of all web-documents including a particular lexicon (e.g., the name of a particular vegetable) during a peculiar period. $search_t^i$ stands for 'search', which is the total number used for searching a particular lexicon during a particular period. unb_t^i stands for 'unbalanced', implying TF-IDF suggested by Salton and McGill (1983). pec_t^i stands for 'peculiar', indicating an index for peculiar lexicon which doesn't appear at ordinary time. So, if a peculiar lexicon appears during a certain periods, it could be a lexicon people are suddenly interested in (Sebastiani, 2002). Finally, $link_t^i$ stands for 'link', and means an index for measuring the importance of linkages among lexicons (Freeman, 1979).

4. Results

Based on time-series data from 2007/07 to 2016/03, we predict each vegetable price for three months from 2016/04 to 2016/06 across BSTS models (BSTS I ~ BSTS III). In order to measure how well each BSTS model predicts vegetable price, we use the following mean absolute percentage error (MAPE) as the criteria of prediction performance.

$$MAPE^i = \frac{1}{n^i} \sum_{t=1}^{n^i} \left| \frac{\text{ACTUAL}_t^i - \text{PREDICT}_t^i}{\text{ACTUAL}_t^i} \right| \tag{17}$$

, where ACTUAL_t^i and PREDICT_t^i are actual price and predicted price for i-vegetable at time period t. Dividing the whole period for prediction periods into the in-sample performance period and the out-of-sample performance period, we apply MAPE only to the in-sample performance period. Whereas, future prices from 2016/04 to 2016/06 are predicted only in the out-of-sample performance period. Those performance periods could be set up differently according to the properties of vegetables. Results are shown in <Table 4> ~ <Table 6> across BSTS models with calculated MAPEs.

<Table 4 ~ Table 6, here>

For garlic, prediction power is higher as atypical indexes are introduced moving from BSTS-I to BSTS-III with lower MAPEs. As for atypical indexes, ‘search’ and ‘unbalance’ indexes are considered in the model.

For onion, the effects of introduction of atypical indexes in BSTS-III are the strongest among three vegetables. As for atypical indexes, ‘unbalance’ and ‘link’ indexes are used. This is interesting result in our paper. There is a popular singer named as ‘Onion’ in Korea, which means the same lexicon could be typed via websites. So, among atypical indexes, some particular indexes are suitable for particular vegetables.

For dried red pepper, the overall results are similar to those of garlic and onion. As for onion, even in BSTS-I and BSTS-II, MAPEs are low, implying that BSTS models are most appropriate for predicting dried red pepper prices.

5. Conclusions

By introducing atypical indexes into the Bayesian structural time series models, we could see that prediction power for vegetable prices are improved. In other words, it can provide better performances in predicting prices to combine recent Big-Data generated atypical web-search data. Especially, it would be valuable if we apply more atypical data into the field of agricultural economics such as food sector, yield, and etc. other than price like in our paper.

Results show as follows: first, the introduction of atypical index obtained from atypical web-search data can improve price prediction power. Second, the improvement across BSTS models could be different by the kind of vegetables. Third, different types of atypical indexes can be used by reflecting the properties of vegetables due to complicate meaning of lexicons like the case of ‘onion’ in Korea.

References

Bollen, J., H. Mao, and X. Zeng, 2011, "Twitter Mood Predicts the Stock Market," *Journal of Computational Science*, 2(1): 1-8.

Choi, H. and H. Varian, 2012, "Predicting the Present with Google Trends," *Economic Record*, 88(1): 2-9.

DeGroot, M. H., 2004, *Optimal Statistical Decisions*, John Wiley & Sons.

Durbin, J. and S. J. Koopman, 2002, "A Simple and Efficient Simulation Smoother for State Space Time Series Analysis," *Biometrika*, 89, 603-616.

Freeman, L. C., 1979, "Centrality in Social Networks Conceptual Clarification," *Social Networks*, 1: 215-239.

Salton, G. and M. J. McGill, 1983, *Introduction to Modern Information Retrieval*, McGraw Hill Book Co., New York.

Harvey, A. C. and N. Shephard, 1993, "Structural Time Series Models," *Handbook of Statistics*, Vol. 11, Elsevier Science Publishers.

Harvey, A. C. and S. Peters, 1990, "Estimation Procedure for Structural Time Series Models," *Journal of Forecasting*, Vol. 9, 89-108.

Koop, G., 2003, *Bayesian Econometrics*, *Chapter 8. Introduction to Time Series: State Space Models*, Wiley, U. K.

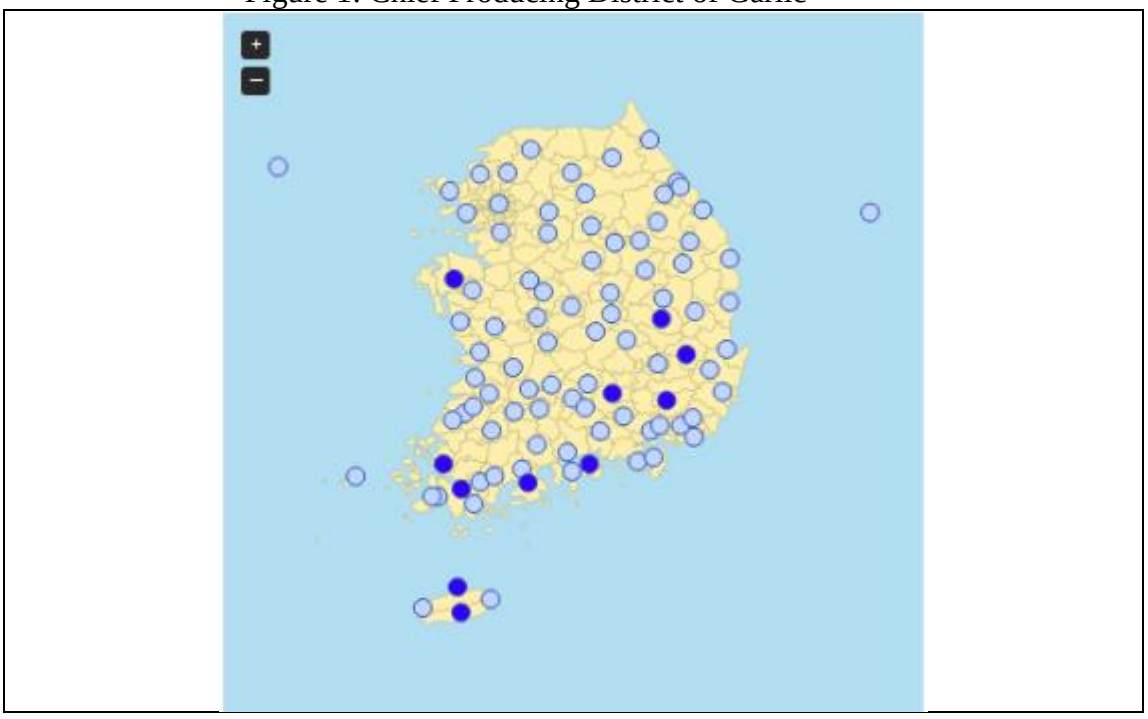
Scott, S. and H. Varian, 2013a, "Bayesian Variable Selection for Nowcasting Economic Time Series," NBER Working Paper 19567.

Scott, S. and H. Varian, 2013b, "Predicting the Present with Bayesian Structural Time Series," Available at SSRN: http://ssrn.com/abstract=2304426.

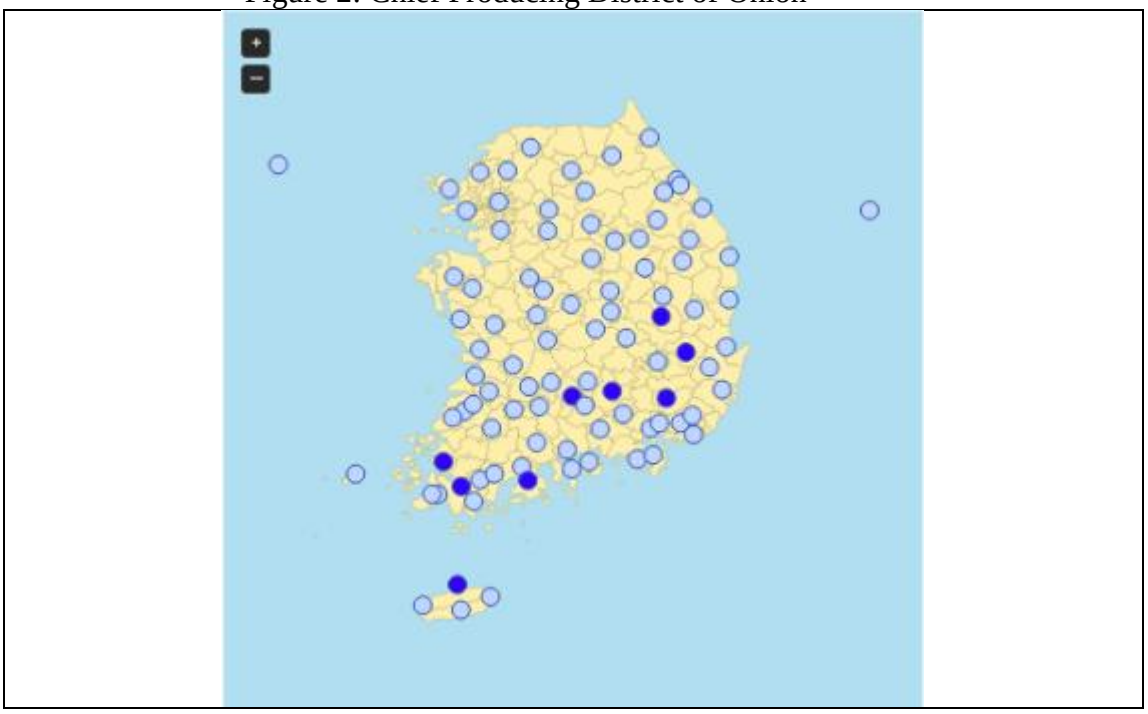
Sebastiani, F., 2002, "Machine Learning in Automated Text Categorization," *ACM Computing Surveys*, 34(1): 1-47.

Witten, I. H., E. Frank, and M. A. Hall, 2011, *Data Mining: Practical Machine Learning Tools and Techniques, 3rd edition*, Burlington, MA: Morgan Kaufmann.

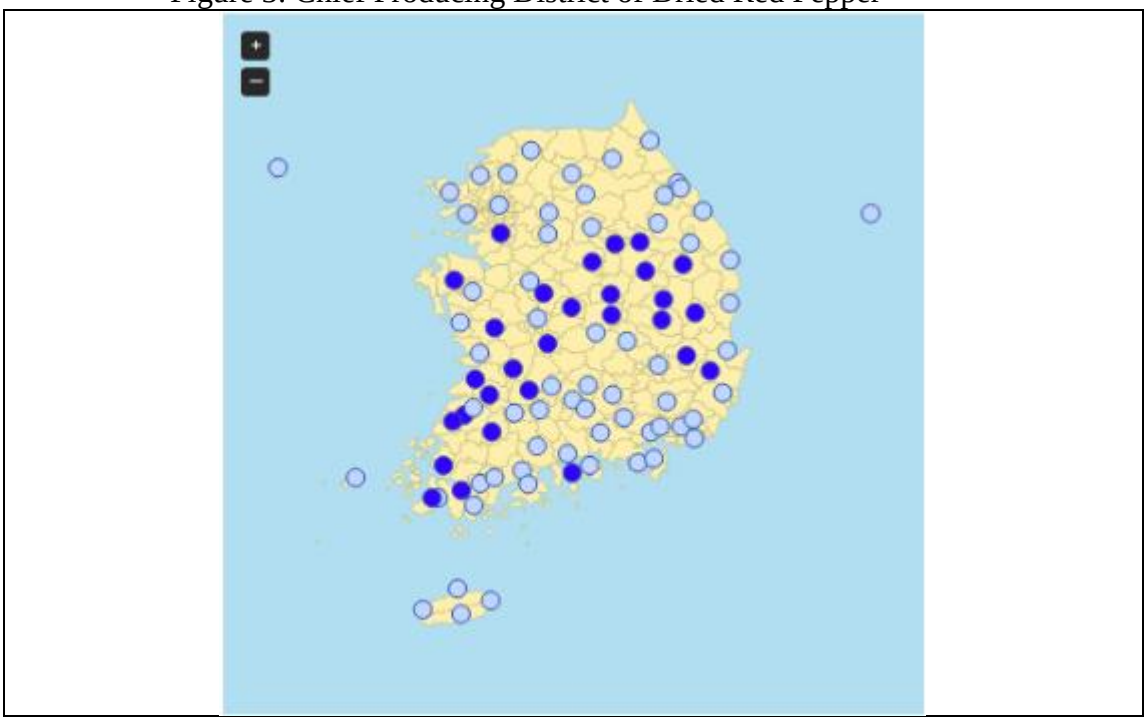
<Figure 1: Chief Producing District of Garlic>



<Figure 2: Chief Producing District of Onion>



<Figure 3: Chief Producing District of Dried Red Pepper>



<Table 1: Descriptive Statistics for Garlic>

	Obs.	Mean	s. d.	Min	Max
Price (KRW/kg)	196	2799.793	1202.675	1348.8	6337.369
Quantity (kg)	196	45648.65	65892.33	703.7037	382382
Temperature (℃)	196	13.64889	8.59474	-2.26129	27.79707
Minimum Temperature (℃)	196	8.957269	9.104615	-7.05455	24.37595
Precipitation (mm)	196	6.237111	4.464317	0.387669	20.64492
Wind speed (m/s)	196	5.005904	0.648675	3.815455	7.347879
Sunshine (Hr)	196	5.852134	1.173345	2.981515	9.002933

<Table 2: Descriptive Statistics for Onion>

	Obs.	Mean	s. d.	Min	Max
Price (KRW/kg)	196	770.2906	365.8856	257.5	2487.727
Quantity (kg)	196	614301	114086.5	422478.5	994000
Temperature (℃)	196	13.36595	8.745036	-2.8914	27.64409
Minimum Temperature (℃)	196	8.38575	9.268761	-7.92616	24.17276
Precipitation (mm)	196	5.889341	4.338769	0.309028	23.2573
Wind speed (m/s)	196	5.00383	0.715785	3.622984	7.621667
Sunshine (Hr)	196	5.824877	1.219705	2.782222	9.054584

<Table 3: Descriptive Statistics for Dried Red Pepper>

	Obs.	Mean	s. d.	Min	Max
Price (KRW/kg)	196	10773.88	4202.027	4941.115	25766.8
Quantity (kg)	196	8474.675	5470.844	0	30363.64
Temperature (℃)	196	12.13733	9.406193	-5.75539	26.79989
Minimum Temperature (℃)	196	7.058058	9.83829	-11.2632	23.60119
Precipitation (mm)	196	5.110349	4.31276	0.343928	21.38899
Wind speed (m/s)	196	4.772991	0.630886	3.680864	7.1424
Sunshine (Hr)	196	5.958112	1.271805	2.493	9.350933

<Table 4: Price Prediction for Garlic across BSTS Models, 2016/04 ~ 2016/06 >

(Unit: KRW/kg)

BSTS Model / Month	BSTS I	BSTS II	BSTS III
	pure time series	w/ climate factors	w/ climate factors & atypical indexes
2015/9	4,584	4,584	4,584
2015/10	5,190	5,190	5,190
2015/11	5,570	5,570	5,570
2015/12	5,716	5,716	5,716
2016/1	5,862	5,862	5,862
2016/2	6,030	6,030	6,030
2016/3	5,781	5,781	5,781
2016/4	4,631	4,831	5,879
2016/5	4,709	4,966	6,052
2016/6	4,774	5,121	6,242
MAPE (2015.09.~2016.03.)	0.2296	0.2280	0.0712

<Table 5: Price Prediction for Onion across BSTS Models, 2016/04 ~ 2016/06 >

(Unit: KRW/kg)

BSTS Model / Month	BSTS I	BSTS II	BSTS III
	pure time series	w/ climate factors	w/ climate factors & atypical indexes
2015/9	1,400	1,400	1,400
2015/10	1,417	1,417	1,417
2015/11	1,594	1,594	1,594
2015/12	1,717	1,717	1,717
2016/1	1,673	1,673	1,673
2016/2	1,632	1,632	1,632
2016/3	1,608	1,608	1,608
2016/4	833	1,172	1,626
2016/5	822	1,261	1,634
2016/6	832	1,341	1,681
MAPE (2015.09.~2016.03.)	0.4806	0.3486	0.0634

<Table 6: Price Prediction for Dried Red Pepper across BSTS Models, 2016/04 ~ 2016/06 >

(Unit: KRW/kg)

BSTS Model / Month	BSTS I	BSTS II	BSTS III
	pure time series	w/ climate factors	w/ climate factors & atypical indexes
2015/4	13,667	13,667	13,667
2015/5	13,667	13,667	13,667
2015/6	13,667	13,667	13,667
2015/7	13,667	13,667	13,667
2015/8	13,670	13,670	13,670
2015/9	13,883	13,883	13,883
2015/10	13,687	13,687	13,687
2015/11	13,497	13,497	13,497
2015/12	13,332	13,332	13,332
2016/1	13,013	13,013	13,013
2016/2	13,000	13,000	13,000
2016/3	12,891	12,891	12,891
2016/4	11,478	15,601	12,653
2016/5	11,561	15,511	12,873
2016/6	11,508	15,677	12,569
MAPE (2015.04.~2016.03.)	0.1317	0.0763	0.0158