Journal of Nuts 3(2):45-52, 2012
ISSN 2008-9937

Seasonal Autoregressive Models for Estimating the Probability of Frost in Rafsanjan

A. Hosseini[1*], *M.S. FallahNezhad*[1], *Y. ZareMehrjardi*[1], *R. Hosseini*[2]

[1]Department of industrial engineering, Yazd University, Yazd, Iran
[2]Division of Biostatistics, University of Southern California, USA

Abstract: This work develops a statistical model to assess the frost risk in Rafsanjan, one of the largest pistachio production regions in the world. These models can be used to estimate the probability that a frost happens in a given time-period during the year; a frost happens after 10 warm days in the growing season. These probability estimates then can be used for: (1) assessing the agroclimate risk of investing in this industry; (2) pricing of weather derivatives. Autoregressive models with time-varying coefficients and different lags are compared using AIC/BIC/AICc and cross validation criterions. The optimal model is an AR (1) with both intercept and the “auto-regressive coefficients” vary with time. The long-term trends are also accounted for and estimated from data. The optimal models are then used to simulate future weather from which the probabilities of appropriate hazard events are estimated.

Keywords: Pistachio, Frost, Weather derivative, Minimum temperature, Time-varying autoregressive coefficients

INTRODUCTION

Rafsanjan County in the north of Kerman Province in Iran is a region with the largest pistachio production in the world and most of the region's economy relies on pistachio production [1].

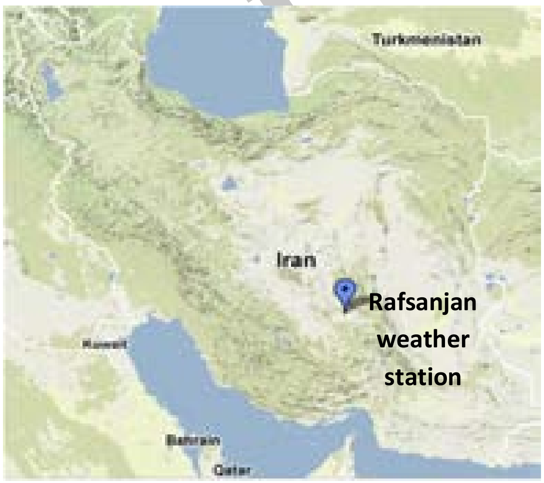

Fig 1.Map and geographic location of Rafsanjan.Latitude 30 25 N, Longitude 55 54 E, Elevation 1580.9 MET. The above map was created using google maps.

In the recent years the most important risk factor for pistachio producers and industry (e.g. farmers, distributers) has been frosts that have destroyed a large proportion of the yield. Therefore methods that can estimate the probability of such events are useful. In particular such methods can: (1) assess the agroclimate risk of investing in this industry; (2) be used in pricing of weather derivatives. In fact, weather derivatives, which may be created as part of a risk management program, can be written in terms of the attainment or non-attainment of specific target-values stipulated in the contract. Temperature-related trades account for 80% of the transactions among all weather derivatives [2]. Most of the work in this area has focused on HDD/CDD (heating degree days/cooling degree days) (e.g. [3, 4]). In this paper we focus on the occurrence of frosts, an issue recently considered in [5], for agricultural crops in Canada.

***Corresponding author:** A. Hosseini, Department of industrial engineering, Yazd University, Iran
E-mail: alirezahosseini65@gmail.com

The models developed in this paper can be applied to estimate the probability that a given period is frost-free; the probability that a given day is the start of a long frost-free period; the distribution of the length of the frost-free period and so on. The same model can be used to compute the probability that a given day of the year is the beginning of the growing season (the first day that the mean temperature is higher than 5 degrees for 5 consequent days) as well as the length of the growing season which are important for agricultural applications. For example in this study we estimate the probability of a useful event: "the minimum temperature goes below zero at least one day in the period March 27th-April 20th". This is an important event because it coincides with the general flowering time of pistachio trees.

Laboratory studies in Rafsanjan region show that temperatures below -2 degrees Celsius damage to pistachio buds. Also temperatures below +2 degrees Celsius damages open flowers [6]. In this study we considered zero degree Celsius as a critical point, because in flowering period some of the crops are buds and some of them are open flowers. However the same model we developed here can be applied to other thresholds such as +2 and -2 degrees Celsius.

Throughout this paper, temperature is measured in degrees Celsius. Let us denote the minimum temperature series by {Y (t)}, t = 0, 1, 2… where t denotes time. We let F to be the investor's defined frost which we take it to be zero in this work. Then we can define the binary frost process:

$$Y_F(t) = \begin{cases} 1 & Y(t) \le F(\deg\mathrm{C}) \\ F & Y(t) > F(\deg\mathrm{C}). \end{cases} \qquad (1)$$

In order to study frosts, we can use these approaches among others: (a) Fit the continuous-valued Markov model to the Y (t) chain; (b) Fit a binary Markov model to the Y_F (t) chain. Hosseini et al. [5] suggest using binary Markov models to avoid assumptions regarding the distribution of temperature and gain robustness for modeling frosts in Alberta, Canada. They show time-varying high-order Markov models with complex seasonal structure are needed and therefore their computations become challenging. Here we investigate Method (a) in fitting such chains and calculating the probabilities of frost events. The advantages of Method (a) are: (1) the fitting can be done with standard packages such as R with less computational problems; (2) only this method can estimate the probability of complicated events. One such complicated event is: "the temperature in March-April is above 5 (deg C) for at least 3 consecutive days and is below zero after". A comparison of the two methods in terms of estimation when they are both applicable is left to future research.

It is clear the temperature series away from the equator is non-homogenous in time because of the seasonal effects during the year (resulted from the relative location of Sun and the Earth and the tilt of the Earth's axis relative to the plane of revolution). Therefore time series models that allow for a seasonally-varying "mean structure" are obviously needed. However it is less trivial that the autocorrelation structure also varies with season. [7, 5] show that the temperature autocorrelation varies with season (in Los Angeles and Alberta respectively). In [7] the authors found that the autocorrelation values in summer and winter are different, therefore they propose to fit two separate models for winter and summer. But they expressed that: "In any event, we stress that this is only a temporary solution; a more satisfactory one would be to fit the entire data series using seasonally varying parameters". In this paper we solve this problem by allowing the coefficients of the autoregressive models to vary with season and showing how such models can be fit in standard statistical packages.

STATISTICAL MODELS

The data in this study are daily minimum temperature values collected at Rafsanjan weather station from 1992 to 2010. Currently, we do not have access to more data from other stations in the area but we hope to acquire those data for future studies to offer more local predictions. In order to model frost occurrences, we introduce statistical models for minimum daily temperature in Rafsanjan. Several features of the temperature process should be considered in modeling:

1. Seasonal trends over time: the temperature process is driven mostly by the sun energy. This energy is dependent on the position of earth relative to sun which goes through a periodic cycle through the year.

2. Long-term trends: Other than seasonal patterns in temperature, long-term climate patterns can be present in the temperature process, for example due to greenhouse gas emissions (climate change) or long-term natural climate shifts as a result of large volcanic activities and so on.

3. Dependence in time: seasonal and long-term climate shifts alone cannot explain the variation in the temperature process in short time-scales. The weather is also influenced by short-time weather regimes that last for a few days to couple of weeks. This causes time dependence in the weather data which can be modeled by relating the minimum temperature of a given day to a few days previous to that. Markov chains (or high-order Markov chains) are the natural statistical framework to model such dependence using the conditional distributions.

Let {Y (t)}, t = 0, 1, 2… T denote the daily minimum temperature process in centigrade, where t denotes the day starting from March 1st 1992 to December 28th 2010. Here we consider autoregressive models with a seasonal component and various lags:

$$Y(t) = \mu(t) + \epsilon(t), \qquad (2)$$

$$\mu(t) = a_0(t) + \sum_{i=1}^{r} a_i Y(t-i) \qquad (3)$$

Where μ(t)=E{Y(t)|Y(t-1),Y(t-2),…} is the conditional mean of minimum temperature at time t; ε(t) are independent identically distributed normal errors ε(t) ~ N(0,σ^2); a_0(t) is the fixed trend coefficient; a_1, a_2,…,a_r are autoregressive coefficients. We allow a_0(t) to include both seasonal and long-term effects by using a Fourier series with period, $\omega = \frac{2\pi}{366}$, and a quadratic trend:

$$a_0(t) = \{a_0 + \sum_{j=1}^{k} a_j \cos(j\omega t) + \beta_j \sin(j\omega t)\}$$
$$+ \{\gamma_1 t + \gamma_2 t^2\}, \qquad (4)$$

In fact in this paper we extend the above model by allowing a_1(t),...,a_r(t) also vary with season by using Fourier series. This was to allow for the autocorrelation to vary with season as suggested in [7].

To fit our models, we use the partial likelihood maximization. By the way of an introduction, we would note that "generalized linear models" were developed to extend ordinary linear regression to the case that the response is not normal. However, that extension required the assumption of independently observed responses. The notion of partial likelihood was introduced to generalize these ideas to time series where the data are dependent. The following definition from [8], gives a more precise description.

Definition: Let F_t ,t=1,2,… be an increasing sequence of$\sigma -$fields, F_0,F_1,F_2, … and let Y_1,Y_2,… be a sequence of random variables such that Y_t is F_t-measurable. Denote the density of $Y_t$given F_t ,by F_t(Y_t;θ), where $\theta \in \mathbb{R}^p$is a fixed parameter. The partial likelihood (*PL*) is defined by:

$$PL(\theta; y_1, \dots, y_N) = \prod_{t=1}^{N} f_t(y_t\,; \theta). \qquad (5)$$

The reader unfamiliar with$\sigma -$fieldsnotion can think of F_t as the information available to us up to time t.

As an example, suppose Y_t represents the 0-1 frost day process in Rafsanjan. We can define $\mathcal{F}_t = \sigma\{Y_{t-1}, Y_{t-2}, \ldots\}$. In this case, we are assuming the information available to us is the value of the process on each of the previous days. If moreover we assume that Y_t is a fixed-coefficient 2nd-order stationary Markov chain, then $P(Y_t|\, \mathcal{F}_t) = P(Y_t|Y_{t-1}, Y_{t-2})$.

We define $Z_{t-1} = (1, Y_{t-1}, Y_{t-2}, Y_{t-1}, Y_{t-2})$ to be the covariate process in the sense that$Y_t|Z_{t-1} = (\theta Z_{t-1}) + \epsilon(t)$,which is a linear form. Other useful covariate processes can be considered. For example $Z_{t-1} = (1, Y_{t-1}, \cos(\omega t), \sin(\omega t))$corresponds to a non-stationary 1st-order Markov chain.

For any such Z_{t-1}, if we assume $\epsilon(t) \sim N\big(0, \sigma^2(t)\big)$, by definition the log partial-likelihood is equal to:

$$\sum_{t=1}^{N} \log P(Y_t|Z_{t-1}) =$$

$$\sum_{1 \le t \le N} \log\{\frac{1}{\sqrt{2\pi}\sigma(t)} \exp[\frac{-(y_t - \mu_t)^2}{2\sigma(t)^2}]\} \qquad (6)$$

$$\sum_{1 \le t \le N} -\log\{\sqrt{2\pi}\} - \log\{\sigma(t)\} + \frac{-(y_t - \mu_t)^2}{2\sigma(t)^2}$$

Where $\mu_t = E(Y_t|Z_{t-1}) = \theta Z_{t-1}$. In this study we assume $\sigma(t)$ is fixed over time. The vector θ that maximizes the above equation is called the maximum partial likelihood (MPLE); [9] showed its consistency, asymptotic normality and efficiency (under certain regularity conditions).

Such models can be fit easily using standard packages to perform statistical analysis such as R, SAS, etc. The fits can be done easily since the above model can be viewed as a linear autoregressive model. In this paper we used R which is a free and powerful tool for analysis of data.

STATISTICAL MODEL SELECTION

In the above we introduced several autoregressive models of: (1) various lags; (2) various seasonal complexity (number of Fourier terms); (3) various long-term trends. Therefore we need to use some criteria to select an optimal model. The problem of model selection is an important one in statistical theory and application. Various criteria are suggested in the literature for example: AIC in [10]; BIC in [11] and AICc in [12]. Denote the likelihood of the data by L (in this paper the "partial likelihood"), the number of covariates by p and the sample size by n. Then we have:

$$AIC = 2p - 2\ln(L), \qquad (7)$$

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1}, \qquad (8)$$

$$BIC = p\log(n) - 2\ln(L). \qquad (9)$$

Since n in our data is large compared to k, AIC and AICc are very close. In order to perform model selection we used the following algorithm using each of the above criteria (for example AIC).

Model selection algorithm:

Step 1: Finding seasonal model for the fixed coefficient (a_0 (t)): For n=1,2,... in the Fourier series (to model a_0(t)) calculate AIC starting from n=1 until AIC does not decrease anymore.

Step 2: Finding long-term model for the fixed coefficient (a_0 (t)): To the best model found in Step 1 add long-term covariates t, t^2,... until AIC does not decrease.

Step 3: To the best model found in Step 2 add Y(t-1),Y(t-2),Y(t-3),... with fixed coefficients over time until AIC does not decrease any further.

Step 4: To the model found in Step 3 add seasonally-varying autoregressive covariates such as Y(t-1) sin(ωt),Y(t-1)cos(ωt),Y(t-1)sin(2ωt),Y(t-1)cos(2ωt) ,..., one by one until AIC does not decrease any further.

The reason we perform the model selection in steps rather than comparing all possible models is because the number of models to be compared increases exponentially with the number of covariates (explanatory variables such as Fourier series terms, Y1 etc). For example if we use 40 covariates, then we will have to compare $2^{40} \approx 10^{12}$ models which would be computationally infeasible.

We have chosen the steps and their order in the above algorithm (model selection) based on our explanatory analysis (and well-known properties of the temperature processes in regions such as Rafsanjan). For example the reason for performing Step 1 at first is: the seasonal effects in temperature explain most of the variation in the temperature series.

When we compared the models using these criteria, AIC and AICc give rise to the same optimal model while BIC picked a simpler model. In Table 1 we have compared these optimal models using cross-validation error and cross-validated correlation. The cross-validation proceeds by: (1) taking an existing data point out; (2) fitting the model; (3) predicting the value of the point we took out (validation). Then the cross-validation error (CVE) is the mean square error of the predictions and the cross-validation correlation (CVR) is the correlation between the predictions and the observed. Table 1 shows that while the CVE and CVR are very close for the two models, the model picked by AIC/AICc slightly outperforms the one picked by BIC and therefore we use that model for estimation. Figure 2 shows the fitted model for minimum temperature in Rafsanjan.

Table 1. We compare the optimal model picked by AIC and AICc (first row) with the optimal model picked by BIC (second row) using cross validation error and cross-validated correlation.

Criterion	Optimal Model: Z_{t-1}	CVE	CVR
AIC and AIC$_c$	$sin(\omega t), cos(\omega t), \ldots, sin(6\omega t), cos(6\omega t), Y_{t-1}, t, t^2$	2.691	0.9458
BIC	$sin(\omega t), cos(\omega t), Y_{t-1}$	2.696	0.9456

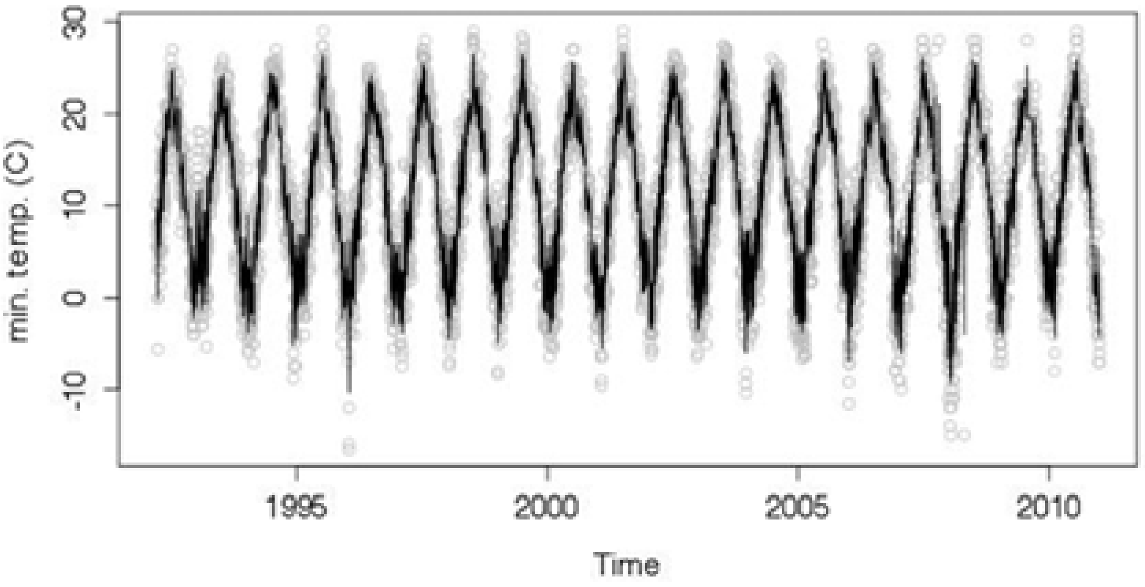

Fig 2.The fitted model for minimum temperature.Circles denote the observed values and the continuous curve denotes the fit from the model

APPLICATIONS IN FROST RISK ASSESSMENT

Previous section found an optimal fit to the data from which estimating the probability of any desired (possibly complex) event is possible by performing multiple simulations. In order to find out the probability of frost in any given day during 2011-2012, we have done 10000 simulations from the model for 2011-2012 and then for each day we have calculated the proportion of frost days (number of frost days divided by 10000). The results are plotted in Figure 1. As we pointed out in the introduction because the flowering time of different varieties of pistachios in Rafsanjan is generally between March 27th and April 20th, it is important to investigate the frost-occurrence during this period which we call the *hazard period*. Figure 2 shows the distribution of the "*number of frost days*" during the hazard period of 2012, where the frequency out of 10000 of any "number of frost days" is plotted. We observe that while it is most likely that no frost occurrs in that period, there is a considerable probability that there is at least one frost. This estimated probability turns out to be about 9 percent which is a plausible number with our experience of pistachio damages caused by frosts in the past 20 years.

In the above we estimated the probability of at least one frost "between March 27th to April 20th", which turned out to be 0.088 (about 9 percent). This is an estimate of the "true probability" and our uncertainty in estimating this probability should be specified. Such an uncertainty is induced from the uncertainty in estimating the regression parameters of the model fitted to the data and therefore depends on the data available and the statistical model. In order to quantify this uncertainty, we calculate a confidence interval for the true probability as follows: (1) We sample a vector of parameters from the estimated variance-covariance matrix of the regression parameters; (2) For each vector of sampled parameters, we estimate the probability by simulations as outlined above. We repeat the above procedure 1000 times and therefore obtain 1000 estimated probabilities. Then to get a confidence interval, we calculate the 0.025th and 0.975 quantities of the 1000 numbers. The 95 percent confidence interval for our data and model turned out to be (0.7, 0.11).

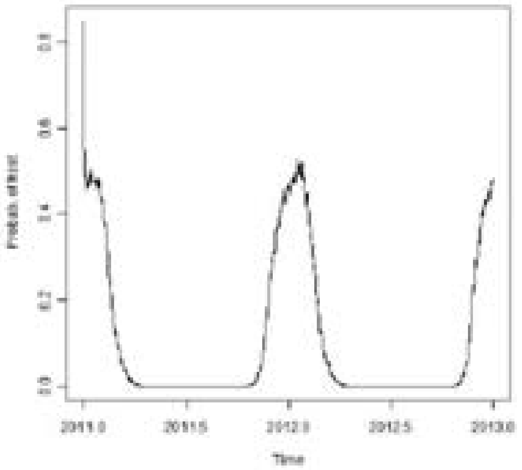

Fig 3. Estimated daily frost probability for 2011-2012 from the model, obtained using 1000 simulations of future Weather. To create this Figure, we estimated the model parameters using available minimum temperature values from 1992 to 2010 and then we simulated the future weather 10000 times to estimate daily probabilities.

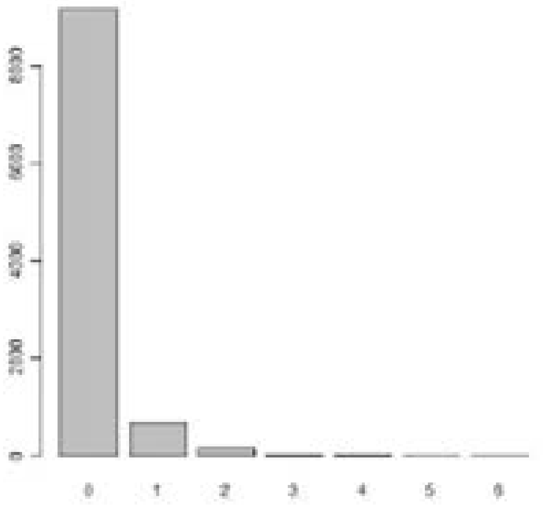

Fig 4.Distribution of frost days during the hazard period (March 27th to April 20th 2012). This is based on 10000 simulations of the future chains for the model fitted to the available data from 1992 to 2010. The probability of at least one frost based on this simulations is 0.0882 which is about 9 percent.

CONCLUSION

This paper developed and compared several statistical models to estimate the probability of hazard frost events for pistachio industry in Rafsanjan. Despite the importance of such risk factors, no systematic estimation of these risks are available in this region as far as we know; this paper is one of the first attempts in developing methods that can assess such risks. Assessing the probabilities of the hazard events are useful in estimating the risk of investing in this industry from production to distribution and exporting. However here we have not investigated other risk factors such as: extremely high temperature during summer; heavy short-time rain during flowering period; slow but long rain during the flowering time. For future studies we plan to acquire the data for precipitation, maximum temperature and developing models that assess these other risk factors.

Another important aspect of assessing the risk is relating the risk factors to the losses in yield or monetary values involved. For this study we relied on expert knowledge (by interviewing farmers and agriculture engineers) to define our hazard period. However if the data for yield per km^2 become available for enough number of years and/or locations, one can develop a statistical model to relate the weather events to the losses in the yield in the same model.

The methodology developed in this work can also be applied to other nuts and crops in various regions of the world. The model selection procedure needs to be repeated to find an appropriate model for different plants and regions. Moreover the critical period should be defined differently depending on the plant and region.

From a statistical modeling point of view two important extensions of these models are subject of our future research. One is the assumption of gaussian errors in the autoregressive models and the other one is the assumption of fixed variance for the errors. Both these assumptions should be assessed and models that do not assume these two assumptions should be compared to the models developed here. However for those extensions standard statistical packages cannot be utilized and we plan to develop packages to estimate such models.

ACKNOWLEDGEMENTS

We are indebted to Mr. Islami for providing the data for this study.

REFERENCES

1. Boshrabadi, HM., Villano, R.A. and Fleming, E., 2007. Production relations and technical inefficiency in pistachio farming systems in Kerman Province of Iran. Forests, Trees and Livelihoods. 17(2):141-155
2. Cao, M. and Wei, J., 2004. Weather derivatives, valuation and market price of weather risk. The Journal of Future Markets. (24)11:1065-1089
3. Richards, T.J., Manfredo, M.R. and Sanders, D.R., 2004. Pricing Weather Derivatives. American Journal of Agricultural Economics, 86(4):1005-1017
4. Benth, F.E., and Benth, J.S., 2007. Volatility of temperature and pricing of weather derivatives. Quantitative Finance, 7(5):553-561
5. Hosseini, R., Le, N. and Zidek, J., 2012. Time-Varying Markov Models for Binary Temperature Series in Agrorisk Management. Journal of Agricultural Biological and Ecological Statistics. 17(2):283-305
6. Sohrabi, N., Hokm-Abadi, H. and Taj-Abadi, A., 2009. Physiology of frostbite in pistachio trees. Pistachio Research Institute, Technical report.
7. Caballero, R., Jewson, S. and Brix, A., 2002. Long memory in surface air temperature: detection, modeling, and application to weather derivative valuation. Journal of Climate Research. 21:127-140

8. Kedem, B. and Fokianos, K., 2002. Regression Models for Time Series Analysis. Wiley Series in Probability and Statistics
9. Wong, W., 1986. Theory of partial likelihood, Annals of Statistics. 14(1):88-123
10. Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19:716-723
11. Schwartz, G., 1978. Estimating the dimension of a model. Annals of Statistics, 6: 461-464
12. Brockwell, P.J. and Davis, R.A., 2009. Time Series: Theory and Methods, 2nd edn, Springer