University of British Columbia
Department of Statistics
Technical Report #257
April 2010

"Model selection for the binary dichotomized temperature processes."

by

Reza Hosseini[1], Nhu D Le[2] and James V Zidek[1]
[1] University of British Columbia
[2] British Columbia Cancer Agency & University of British Columbia

ABSTRACT

This paper proposes using rth-order categorical Markov chains to study the occurrence of extremely high (above a threshold) and low temperatures (below zero). Several stationary and non-stationary higher order Markov models are proposed and compared using BIC. Partial likelihood theory is used to estimate the parameters of these models. The models are then used to build confidence intervals for the probability of a frost-free period in Medicine Hat.

Keywords: binary Markov processes; temperature model; frost; Markov model selection

1 Introduction

This paper develops and demonstrates use of rth-order categorical Markov chains theory developed in [5] to find models for extreme temperature events. A fundamental premise is that temperature itself, which could be handled by standard Gaussian space-time models, is not of specific interest. Instead its dichotomized values play the central role. For example in agroclimate risk analysis and management, the genesis of this paper, any temperature below zero destroys crops. Likewise weather derivatives, which may be created as part of a risk management insurance problem will be set on the attainment or not of a specific target-value stipulated in the contract. Here we consider both low and high temperatures.

The paper develops a modeling strategy that can be used for such things as calculating the likelihood of a long sequence of high temperature days that can, coming at the wrong time of the growing season, reduce crop yield. The same approach can be used for other climatological events, and in fact it was for precipitation in Chapter 4 of [2]. Likelihoods that can be calculated with this approach can play a role in setting crop insurance premiums and managing irrigation programs, which are attaining increasing importance as the climate changes.

The models used here are an extension of the logistic regression for the independent data to dependent case. [5] investigates the estimation of the coefficients of rth order Markov chains with seasonal terms (non-homogenous) using partial likelihood and picking the model using BIC. It shows that the partial likelihood performs very well in picking the true model, the partial likelihood estimates are close to the true values and the distribution of the parameter estimates are close to normal distribution.

Throughout this paper, temperature is measured in degrees centigrade. We call a day with minimum temperature (mt) less than zero "extremely cold" and denote it by e. Thus:

$$e(t) = \begin{cases} 1 & mt(t) \leq 0 \text{ (deg C)} \\ 0 & mt(t) > 0 \text{ (deg C)} \end{cases}.$$

Taking 0 (deg C) to be the cut-off for low temperature seems reasonable in the absence of any other considerations, since it is the usual definition of a frost. In agriculture, where most plants contain a lot of water this can be be considered as an important cut-off. No seemingly natural cut-off like that for minimum temperature exists for extremely high temperature.

Obviously that cut-off will depend on the purpose of the model. In farming, the various crops have different tolerances to hot or cold weather, depending in part on soil conditions. Clearly the definition of extreme may need to depend on the time of the year or location.

Ralph Wright (personal communication) from the Alberta Agriculture Food and Rural Development (AAFRD) made these points more concretely when he said about droughts that: "Drought is really defined by the impact that the moisture deficit has on a specific use or uses. Its definition can vary both with time of year and from place-to-place. Drought can be short-term or long-term. For example, one month of hot dry weather can significantly reduce crop yields, despite the fact that normal amounts of precipitation have been received over the past year. On the other hand, crops may do fine in dry weather conditions if precipitation has been received in a timely manner and temperatures have been favorable. However under the same conditions, a dam operator in the same area may have severe shortages in the reservoir and declare drought like conditions (e.g. with low winter snow-fall and poor spring run-off). You will need to define your drought based on whom or what is being impacted by the water shortage."

Since we do not have any standard definition of an extremely hot day, we use the data to find a plausible choice for the purpose of describing our modeling approach, the central feature of this paper. For that purpose we rely on quantiles since they have long been used to characterize extreme events. In our application, we pick somewhat arbitrarily, the global spatial/temporal 95th percentile of $q = 27$ (deg C) to dichotomize the data and define a binary process of (hot)/(not hot) for temperature. This value is calculated using data from 25 stations over Alberta, which had daily maximum temperature (MT) data from 1940 to 2004. That choice could be made more incisively once specific objectives have been specified in a specific context. That percentile turns out to be . Moreover, we chose a global quantile assuming the definition of a hot day should be the same over the province and across the years. Then we define the binary process of extremely hot temperature as:

$$E(t) = \begin{cases} 1 & MT(t) \geq q \\ 0 & MT(t) < q \end{cases},$$

where $q = 27$ (deg C) here.

In order to study extreme events (e.g. for MT) three approaches are possible:

1. Model the whole daily MT process and use that to infer the extremes. For MT, we have shown that a Gaussian distribution fits the daily values well. However, in the tails, usually of paramount concern, the fit does not do well as shown in the qq-plots in [4]. Another difficulty with this approach is picking a covariance function to model the covariance over time. Also in [2], Hosseini showed that even though two distributions are very close in terms of overall quantile distance, they might not be very close in terms of tail quantile distance. This shows in order to study extremes (for example extremely hot temperature) if we use a good overall fit, our results might not be reliable.

2. Use a specified threshold and model the values exceeding the threshold. This approach has several drawbacks. Firstly we cannot answer the question of how

often or in what periods of the year the extremes happen. This is because we model the actual extreme values and ignore the non-extreme values. Secondly, strong assumption of independence is needed for this method. Thirdly we need to pick the threshold high enough to make the model reasonable. This might not be an optimal threshold from a practical point of view.

3. Based on a real problem, use a threshold to define a new binary process of (extreme)/(not extreme) values and then model that binary process. This is the method we use and it does not have the issues mentioned in 1 and 2 because the threshold is not taken to satisfy some statistical property and we make few assumptions about the binary chain. In any case, we introduce a new method to investigate this chain.

2 rth-order Markov models for extreme minimum temperatures

This section looks for appropriate models for the binary process $e(t)$ of cold/not cold temperature days. This is a binary process and the Categorical Expansion Theorem [2] gives the form of all such rth-order Markov chains. Here we give an example.

Example 2.1 *For the stationary binary (0-1) Markov chain of order $r = 3$ and $t \geq 3$ and a fixed transformation $g : \mathbb{R} \to \mathbb{R}^+$*
There exist unique α parameters:

$$g_t = g^{-1}\{\frac{P(X_t = 1|X_{t-1} = x_{t-1}, \cdots, X_0 = x_0)}{P(X_t = 0|X_{t-1} = x_{t-1}, \cdots, X_0 = x_0)}\}$$

$$= \alpha_0 + \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \alpha_3 x_{t-3}$$

$$+\alpha_{12} x_{t-1} x_{t-2} + \alpha_{23} x_{t-2} x_{t-3} + \alpha_{13} x_{t-1} x_{t-3}$$

$$+\alpha_{123} x_{t-1} x_{t-2} x_{t-3}.$$

Conversely every collection of arbitrary α's corresponds to a unique 3rd-order stationary binary (0-1) Markov chain. If we take $g : \mathbb{R} \to \mathbb{R}^+$ then $g^{-1}(x) = \log(x)$ in the above.

An advantage of this linear form is the ability to estimate the parameters using the partial likelihood as discussed by Kedem and Fokianos in [3]. Another advantage is its capacity to allow other linear terms needed to build non-stationary chains. For example, we can add $cos(\omega t)$ to model seasonality. In the theory of partial likelihood the covariate process is denoted by Z_{t-1}. We denote the 0-1 precipitation process by Y_t and discuss a few models in the following. For notational simplicity we denote Y_{t-i} by Y^i.

Here we also consider other covariates such as the minimum temperature of the previous day, two days ago as well as seasonal covariates (deterministic). The next subsection uses graphical tools and exploratory techniques to investigate the properties the model should have. Then we use the BIC criterion and compare several proposed models. We use partial likelihood techniques to estimate parameters as proposed by Kedem et al. in [3].

2.1 Exploratory analysis for binary extreme minimum temperatures

Here we perform an exploratory analysis of the binary process $e(t)$ using two stations for this purpose, Banff and Medicine Hat which have data from 1895 to 2006. The transition probabilities are computed from the historical data considering years as independent observations. The results are summarized a follows:

- Figures 1 and 2 plot the probability of a freezing day over the course of a year for the Banff and Medicine Hat stations, respectively. A regular seasonal pattern is seen. Medicine Hat seems to have a much longer frost-free period.

- Figures 3 and 4 plot the estimated transition probabilities, $\hat{p}_{01}$ and $\hat{p}_{11}$ for the Banff and Medicine Hat stations. If the chain were a 0th-order Markov chain then these two curves would overlap. This is not the case and Markov chain at least of 1st-order seems necessary. In the $\hat{p}_{01}$ curve for both Banff and Medicine Hat, high fluctuations are seen at the beginning and end of the year which corresponds to the cold season. This is not surprising because there are very few pairs in the data with a freezing day followed by a non-freezing day in a cold season in Alberta.

- In Figure 4, $\hat{p_{11}}$ is missing for a period over the summer. This is because no freezing day is observed over this period in the summer and hence $\hat{p}_{11}$ could not be estimated.

- Figures 5 and 6 give the plots for the 2nd-order transition probabilities. They overlap substantially and hence a 2nd-order Markov chain does not seem to be necessary.

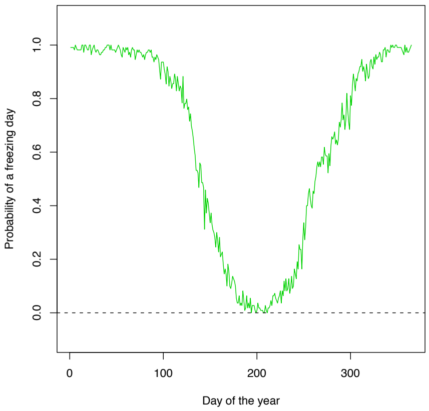

Figure 1: The estimated probability of a freezing day for the Banff site for different days of a year computed using the historical data.

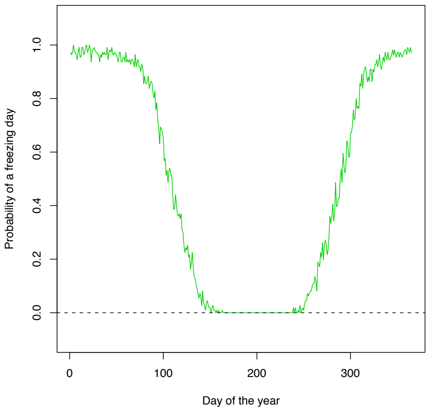

Figure 2: The estimated probability of a freezing day for the Medicine Hat site for different days of a year computed using the historical data.

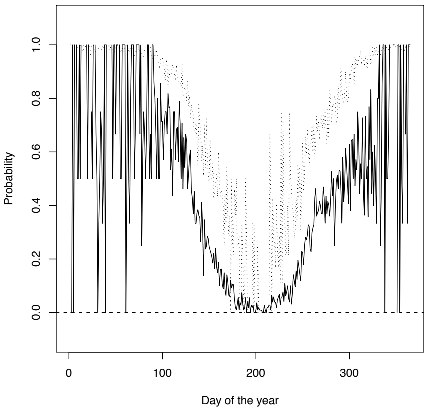

Figure 3: The estimated 1st-order transition probabilities for the 0-1 process of extreme minimum temperatures for the Banff site. The dotted line represents the estimated probability of "$e(t) = 1$ if $e(t-1) = 1$" ($\hat{p_{11}}$) and the dashed, "$e(t) = 1$ if $e(t-1) = 0$" ($\hat{p_{01}}$).

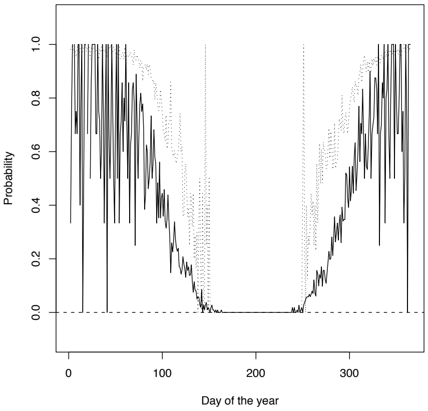

Figure 4: The estimated 1st-order transition probabilities for the 0-1 process of extreme minimum temperatures for the Medicine Hat site. The dotted line represents the estimated probability of "$e(t) = 1$ if $e(t-1) = 1$" ($\hat{p_{11}}$) and the dashed, "$e(t) = 1$ if $e(t-1) = 0$" ($\hat{p_{01}}$).

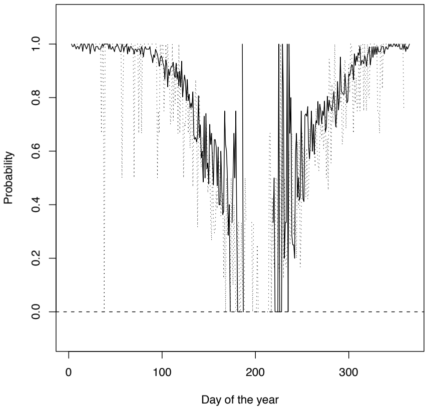

Figure 5: The estimated 2nd-order transition probabilities for the 0-1 process of extreme minimum temperature for the Banff site with $\hat{p}_{111}$ (solid) compared with $\hat{p}_{011}$ (dotted) both calculated from the historical data.

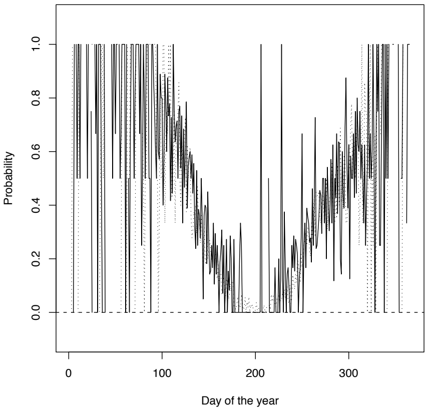

Figure 6: The estimated 2nd-order transition probabilities for the 0-1 process of extreme minimum temperatures for the Banff site with $\hat{p}_{001}$ (solid) compared with $\hat{p}_{101}$ (dotted) calculated from the historical data.

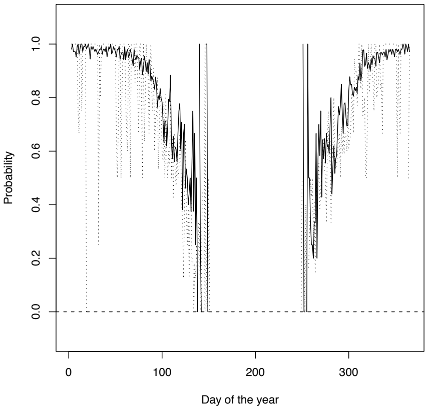

Figure 7: The estimated 2nd-order transition probabilities for the 0-1 process of extreme minimum temperatures for the Medicine Hat site with $\hat{p}_{111}$ (solid) compared with $\hat{p}_{011}$ (dotted) calculated from the historical data.

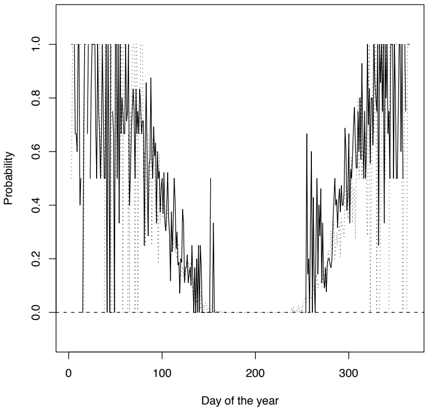

Figure 8: The estimated 2nd-order transition probabilities for the 0-1 process of extreme minimum temperatures for the Medicine Hat site with $\hat{p}_{001}$ (solid) compared with $\hat{p}_{101}$ (dotted) calculated from the historical data.

Model: Z_{t-1}	BIC	parameter estimates
$(1, N^1)$	1251.7	(-2.144, 4.260)
$(1, N^2)$	1166.5	(-2.501, 2.490)
$(1, N^3)$	1142.9	(-2.653, 1.755)
$(1, N^4)$	1121.6	(-2.773, 1.371)
$(1, N^5)$	1111.2	(-2.852, 1.125)
$(1, N^6)$	1093.1	(-2.932, 0.961)
$(1, N^7)$	1087.4	(-2.977, 0.835)
$(1, N^8)$	1081.7	(-3.015, 0.739)
$(1, N^9)$	1077.1	(-3.047, 0.663)
$(1, N^{10})$	1066.5	(-3.089, 0.605)
$(1, N^{11})$	**1056.4**	(-3.130, 0.557)
$(1, N^{12})$	1059.5	(-3.135, 0.511)
$(1, N^{13})$	1062.3	(-3.140, 0.472)
$(1, N^{14})$	1072.8	(-3.126, 0.437)
$(1, N^{15})$	1080.9	(-3.118, 0.406)

Table 1: BIC values for models including N^k for the extreme minimum temperature process $e(t)$ at the Medicine Hat site.

2.2 Model selection for extreme minimum temperature

This section finds models for the extreme minimum temperature process $e(t)$. Here Z_{t-1} denotes the covariate process. We investigate the following predictors:

- $e^k(t) \equiv e(t-k)$. Was it an extremely cold day k days ago?
- $mt^k(t) \equiv mt(t-k)$, the actual minimum temperature k days ago.
- N^k, the number of freezing days during the k previous days.
- SIN, COS, $SIN2$ and $COS2$ which are abbreviations for $\sin(\omega t)$, $\cos(\omega t)$, $\sin(2\omega t)$ and $\cos(2\omega t)$, respectively (with $\omega = \frac{2\pi}{366}$).

Table 1 compares models with a constant and N^k as the covariate process. The optimal model picked by the BIC criterion is the model with the covariates $Z_{t-1} = (1, N^{11})$.

Table 2 compares several models some of which include seasonal terms and continuous variables. The optimal model is $(1, mt^1, COS, SIN)$, which has the temperature of the previous day and seasonal terms. The model $(1, e^1, COS, SIN)$ has a larger BIC but is preferable to all models other than $(1, mt^1, COS, SIN)$ and $(1, mt^1, mt^2, COS, SIN)$. Note that it is not possible to compute the probability of events in the long-term future using $(1, mt^1, COS, SIN)$, since we do not know mt except for perhaps the present time. Hence the optimal applicable model seems to be $(1, e^1, COS, SIN)$.

Model: Z_{t-1}	BIC	parameter estimates
(1)	2539.9	(-0.0251)
$(1, e^1)$	1251.7	(-2.144, 4.260)
$(1, e^2)$	1473.6	(-1.856, 3.683)
$(1, e^1, e^2)$	1157.7	(-2.501, 3.085, 1.896)
$(1, e^1, e^2, e^1e^2)$	1162.4	(-2.586, 3.389, 2.190, -0.593)
$(1, mt^1)$	963.7	(0.109, -0.400)
$(1, mt^1, mt^2)$	954.0	(0.091, -0.329, -0.082)
$(1, COS, SIN)$	984.0	(-0.070, 4.292, 1.324)
$(1, COS, SIN, COS2, SIN2)$	984.2	(-0.502, 4.505, 1.399, -0.464, -0.493)
$(1, COS, SIN, COS2)$	986.7	(-0.258, 4.359, 1.335, -0.353)
$(1, COS, SIN, SIN2)$	984.4	(-0.217, 4.365, 1.360, -0.402)
$(1, mt^1, mt^2, mt^3)$	940.7	(0.062, -0.319, -0.009, -0.094)
$(1, mt^1, mt^2, mt^1mt^2)$	943.4	(0.211, -0.339, -0.084, -0.0091)
$(1, e^1, COS, SIN)$	901.5	(-1.008, 1.840, 3.325, 1.013)
$(1, mt^1, COS, SIN)$	**855.3**	(-0.074, -0.234, 2.394, 0.746)
$(1, mt^1, mt^2, COS, SIN)$	861.9	(-0.076, -0.247, 0.023, 2.504, 0.785)

Table 2: BIC values for several models for the extreme minimum temperature $e(t)$ at the Medicine Hat site.

3 rth-order Markov models for extreme maximum temperatures

This section finds appropriate models for the binary process of extremely hot temperature $E(t)$ as defined above. To define a hot day, we use the 95th percentile of data from 25 stations over Alberta that had daily MT data from 1940 to 2004. The 95th percentile turns out to be $q = 27$ (deg C). Once we used the fast algorithm developed in Chapter 7 of [2] to pick the quantile and once we used an exact method; the algorithm gave us the approximate value $q = 26.7$, which is very close to the exact value.

3.1 Exploratory analysis for extreme maximum temperatures

This section uses explanatory data analysis techniques to study the binary process $E(t)$. Again we use two stations for this purpose, the Banff and Medicine Hat sites that have data from 1895 to 2006. The transition probabilities are computed using the historical data considering years as independent observations. The results are summarized as follows:

- Figures 9 and 10 plot the probabilities of a hot day over the course of a year for the Banff and Medicine Hat stations respectively. A regular seasonal pattern is seen. Medicine Hat seems to have a much longer period of hot days.

- Figures 11 and 12 plot the estimated transition probabilities, $\hat{p}_{01}$ and $\hat{p}_{11}$ for Banff and Medicine Hat. If the chain were a 0th-order Markov chain then these two curves would overlap. This is not the case so Markov chain of at least 1st-order seems necessary. In the $\hat{p}_{01}$ curve for both Banff and Medicine Hat, large fluctuations are seen in the middle of the year, which corresponds to the warm season. This is not surprising because there are very few pairs in the data with a hot day followed by a not–hot day in the warm season in Alberta.

- In Figure 12, $\hat{p_{11}}$ is missing for a period over the cold season. This is because no hot day is observed during this period in the cold season and hence $\hat{p}_{11}$ could not be estimated.

- Figures 13 and 14 give the plots for the 2nd-order transition probabilities. They overlap heavily and hence a 2nd-order Markov chain does not seem to be necessary.

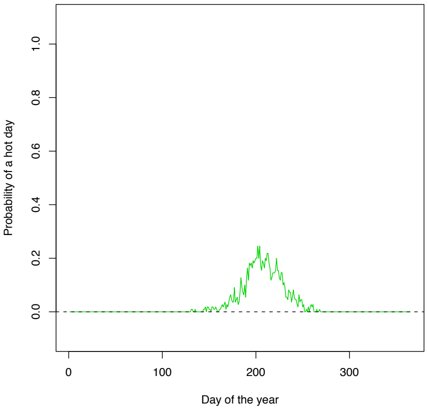

Figure 9: The estimated probability of a hot day (maximum temperature $\geq$ 27 (deg C)) for different days of the year for the Banff site calculated from the historical data.

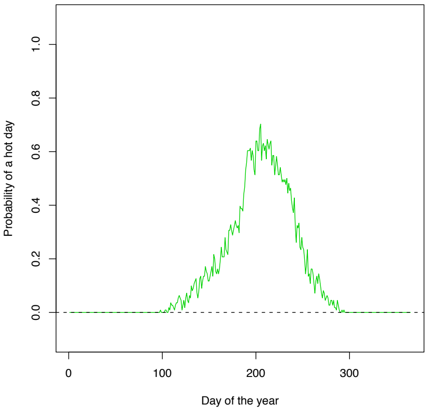

Figure 10: The estimated probability of a hot day (maximum temperature $\geq$ 27 (deg C)) for different days of the year for the Medicine Hat site calculated from the historical data.

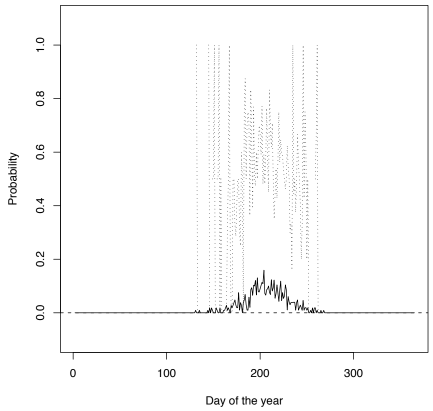

Figure 11: The estimated 1st-order transition probabilities for the binary process of extremely hot temperatures for the Banff site. The dotted line represent the estimated probability of "$E(t) = 1$ if $E(t-1) = 1$" ($\hat{p_{11}}$) and the dashed, "$E(t) = 1$ if $E(t-1) = 0$" ($\hat{p_{01}}$).

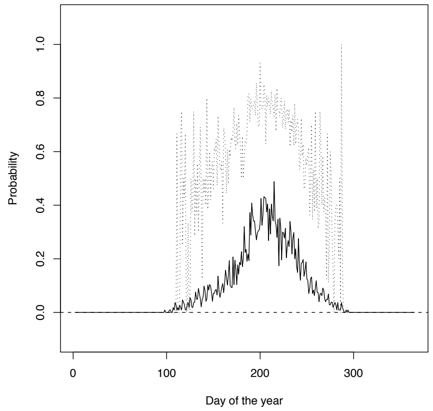

Figure 12: The estimated 1st-order transition probabilities for the binary process of extremely hot temperatures for the Medicine Hat site. The dotted line represents the estimated probability of "$E(t) = 1$ if $E(t-1) = 1$" ($\hat{p_{11}}$) and the dashed, "$E(t) = 1$ if $E(t-1) = 0$" ($\hat{p_{01}}$).

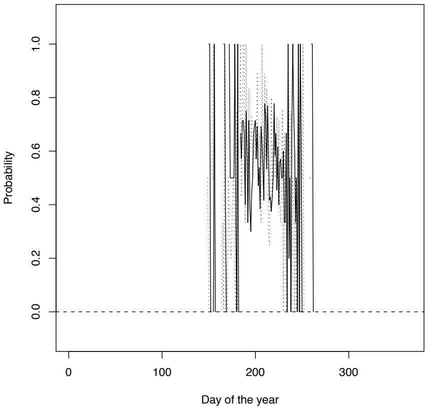

Figure 13: The estimated 2nd-order transition probabilities for the binary process of extremely hot temperatures for the Banff site with $\hat{p}_{111}$ (solid) compared with $\hat{p}_{011}$ (dotted) calculated from the historical data.

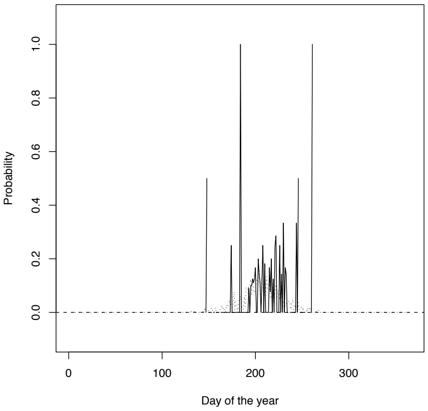

Figure 14: The estimated 2nd-order transition probabilities for the binary process of extremely hot temperatures for the Banff site with $\hat{p}_{001}$ (solid) compared with $\hat{p}_{101}$ (dotted) calculated from the historical data.

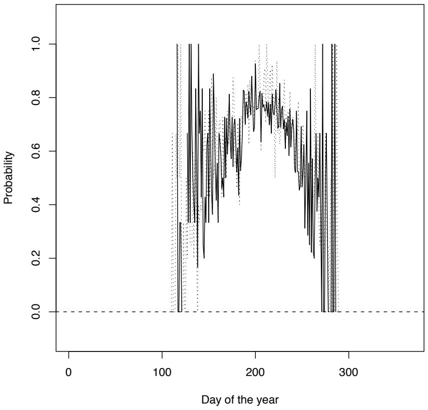

Figure 15: The estimated 2nd-order transition probabilities for the binary process of extremely hot temperatures for the Medicine Hat site with $\hat{p}_{111}$ (solid) compared with $\hat{p}_{011}$ (dotted), calculated from the historical data.

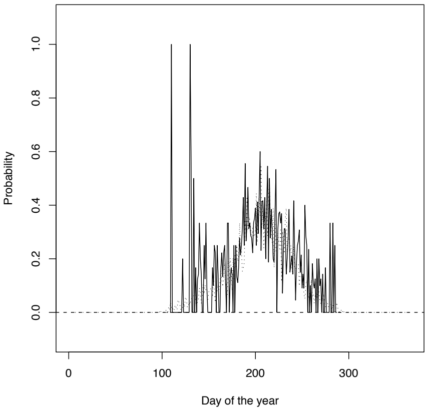

Figure 16: The estimated 2nd-order transition probabilities for the binary process of extremely hot temperatures for the Medicine Hat site with $\hat{p}_{001}$ (solid) compared with $\hat{p}_{101}$ (dotted) calculated from the historical data.

3.2 Model selection for extreme maximum temperature

Here, we use the following abbreviations:

- $E^k(t) = E(t-k)$. Was it an extreme day k days ago?
- $MT^k(t) = MT(t-k)$, the actual maximum temperature k days ago.

- N^k, COS, SIN, COS, $SIN2$ and $COS2$ as previous sections.

Table 3 compares several models containing N^k. The optimal model turns out to be $(1, N^{11})$ which is the same as the result for the extreme minimum temperature process $e(t)$.

Model: Z_{t-1}	BIC	parameter estimates
$(1, N^1)$	955.7	(-2.95, 3.82)
$(1, N^2)$	965.9	(-3.00, 2.16)
$(1, N^3)$	942.5	(-3.11, 1.60)
$(1, N^4)$	921.8	(-3.20, 1.29)
$(1, N^5)$	926.8	(-3.23, 1.05)
$(1, N^6)$	931.6	(-3.24, 0.89)
$(1, N^7)$	932.5	(-3.26, 0.78)
$(1, N^8)$	939.0	(-3.26, 0.69)
$(1, N^9$	931.6	(-3.29, 0.63)
$(1, N^{10})$	925.9	(-3.31, 0.57)
$(1, N^{11})$	**911.7**	(-3.35, 0.49)
$(1, N^{12})$	917.5	(-3.34, 0.46)
$(1, N^{13})$	922.8	(-3.33, 0.42)
$(1, N^{14})$	926.0	(-3.32, 0.39)
$(1, N^{15})$	932.1	(-3.31, 0.37)

Table 3: BIC values for models including N^k for the extremely hot process $E(t)$.

Table 4 compares several models. We observe that major reductions are seen if we use MT^k instead of E^k. The optimal model turns out to be $(1, MT^1, COS, SIN)$ which is combination of seasonal terms and the temperature of the day before.

Model: Z_{t-1}	BIC	parameter estimates
(1)	1520.3	(-1.774)
$(1, E^1)$	955.8	(-2.95, 3.82)
$(1, E^2)$	1170.5	(-2.581, 2.924)
$(1, E^1, E^2)$	941.3	(-3.034, 3.179, 1.099)
$(1, E^1, E^2, E^1E^2)$	929.0	(-3.202, 3.895, 2.137, -1.877)
$(1, MT^1)$	683.8	(-10.040, 0.362)
$(1, MT^1, MT^2)$	689.1	(-10.135, 0.333, 0.034)
$(1, COS, SIN)$	830.8	(-5.484, -5.616, -2.452)
$(1, COS, SIN, COS2, SIN2)$	837.5	(-4.343, -4.255, -0.993, 0.113, 1.016)
$(1, COS, SIN, COS2)$	837.9	(-5.850, -6.231, -2.406, -0.292)
$(1, COS, SIN, SIN2)$	830.0	(-4.481, -4.492, -0.978, 1.011)
$(1, MT^1, MT^2, MT^3)$	669.2	(-10.885, 0.338, -0.061, 0.120)
$(1, MT^1, MT^2, MT^1MT^2)$	681.9	(-21.003, 0.763, 0.452, -0.0162)
$(1, E^1, COS, SIN)$	731.3	(-4.963, 2.005, -4.096, -1.685)
$(1, MT^1, COS, SIN)$	**649.9**	(-10.281, 0.283, -2.829, -1.079)
$(1, MT^1, MT^2, COS, SIN)$	657.3	(-10.109, 0.294, -0.011, -2.609,-1.072)

Table 4: BIC values for several models for the extremely hot process $E(t)$.

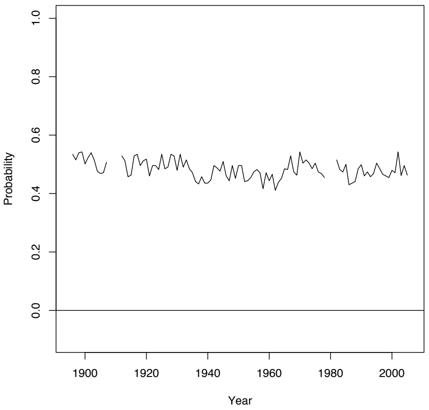

Figure 17: Medicine Hat's estimated mean annual probability of frost calculated from the historical data.

4 Probability of a frost-free period for Medicine Hat

This section shows how the approach developed above can be used in applications. We use the developed methodology to compute two probabilities:

- π_1 : The probability of no frosts in the first week of October at the Medicine Hat site.
- π_2 : The probability of at least 5 days without frost in the first week of October at the Medicine Hat site.

The first day of October is the 275th day of the year in a leap year and the 274th day of the year in a non-leap year. We compute the probabilities for the week between 274th day and 281th day which corresponds to the first week of October in a non-leap year. We prefer this option to computing the probability for the actual first week of October, since this corresponds better to the natural cycles. Of course with a little modification one could compute the probability for the first week of October, for example by introducing a probability of 1/4 for being in a leap year.

Figure 17 plots the probability of a frost for each day of years since 1985. Only years with more than 355 days of data are considered. The figure shows that the probability of a frost is fairly consistent over the years, so we assume a constant probability of frost for all years. Table 5 compares models with various N^k. The optimal model is $(1, N^{11})$. Table 6 includes two seasonal terms as well as N^k. The optimum this time $(1, N^1, COS, SIN)$, showing that in the presence of seasonal terms, the short-term past modeled by N^k is not necessary.

Model: Z_{t-1}	BIC
$(1, N^1)$	5072.2
$(1, N^2)$	4634.8
$(1, N^3)$	4465.9
$(1, N^4)$	4407.4
$(1, N^5)$	4366.0
$(1, N^6)$	4357.4
$(1, N^7)$	4356.2
$(1, N^8)$	4342.6
$(1, N^9)$	4330.5
$(1, N^{10})$	4329.1
$(1, N^{11})$	**4328.4**
$(1, N^{12})$	4332.4
$(1, N^{13})$	4330.8
$(1, N^{14})$	4345.1
$(1, N^{15})$	4362.9

Table 5: BIC values for models including N^k for the extremely cold process $e(t)$ at the Medicine Hat site.

Model: Z_{t-1}	BIC
$(1, N^1, COS, SIN)$	**3601.3**
$(1, N^2, COS, SIN)$	3654.8
$(1, N^3, COS, SIN)$	3693.9
$\vdots$	$\vdots$
$(1, N^{10}, COS, SIN)$	3843.6
$(1, N^{11}, COS, SIN)$	3849.8
$(1, N^{12}, COS, SIN)$	3855.5

Table 6: BIC values for several models including N^k and seasonal terms for the extremely cold process $e(t)$ at the Medicine Hat site.

Model: Z_{t-1}	BIC	parameter estimates
(1)	10122.4	(-0.0858)
$(1, e^1)$	5072.2	(-2.13, 4.18)
$(1, e^1, e^2)$	4598.2	(-2.530, 2.977, 2.00)
$(1, e^1, e^2, e^1e^2)$	4582.8	(-2.65, 3.41, 2.43, -0.855)
$(1, COS, SIN)$	3916.870	(-0.3, 4.301, 1.139)
$(1, COS, SIN, COS2, SIN2)$	3865.6	(-0.746, 4.643, 1.253 -0.550 -0.504)
$(1, e^1, COS, SIN)$	3601.3	(-1.116, 1.760, 3.332, 0.856)
$(1, e^1, COS, SIN, COS2, SIN2)$	**3566.7**	(-1.49, 1.71, 3.65, 0.96, -0.48, -0.42)
$(1, e^1, e^2, COS, SIN)$	3601.6	(-1.22, 1.66, 0.33, 3.19, 0.810)
$(1, e^1, e^2, COS, SIN, COS2, SIN2$ $, COS3, SIN3)$	3571.7	(-1.8, 1.7, 4.4, 1.26, -0.78, -0.74, 0.21, 0.44)
$(1, mt^1, COS, SIN, COS2, SIN2)$	**3356.4**	(-0.66, -0.22, 2.85, 0.73, -0.56, -0.42)

Table 7: BIC values for several models for the extremely cold process $e(t)$ at the Medicine Hat site.

Covariate	Theoretical sd	Experimental sd
1	0.090	0.093
e^1	0.097	0.100
COS	0.125	0.139
SIN	0.060	0.059
$COS2$	0.089	0.094
$SIN2$	0.081	0.077

Table 8: Theoretical and simulation estimated standard deviations for extremely cold process $e(t)$ at the Medicine Hat site.

Table 5 compares various models. The winner is

$$(1, mt^1, COS, SIN, COS2, SIN2).$$

However, it is not possible to compute the desired probabilities using this model since we do not know mt^1 (perhaps except at the start of the chain). Among all other models, the optimal is $(1, e^1, COS, SIN, COS2, SIN2)$ which we use to compute the probabilities.

We compute the standard deviations once using simulations by generating chains from the fitted model with covariates $(1, e^1, COS, SIN, COS2, SIN2)$, and once by computing the partial information matrix, G_N, using partial likelihood theory. The results are given in Table 8. The variance-covariance matrix calculated using partial likelihood theory is given below:

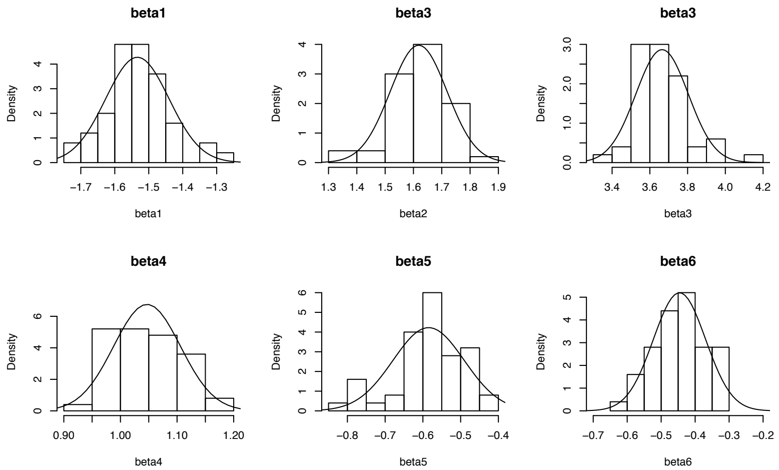

Figure 18: Normal curved fitted to the distribution of 50 samples of the estimated parameters.

$$\begin{pmatrix} 0.0082 & -0.0043 & -0.0038 & -0.0011 & 0.0050 & 0.0030 \\ -0.0043 & 0.0094 & -0.0042 & -0.0013 & 0.0002 & 0.0003 \\ -0.0038 & -0.0042 & 0.0158 & 0.0038 & -0.0052 & -0.0037 \\ -0.0011 & -0.0013 & 0.0038 & 0.0037 & -0.0011 & -0.0017 \\ 0.0050 & 0.0002 & -0.0052 & -0.0011 & 0.0079 & 0.0015 \\ 0.0030 & 0.0003 & -0.0037 & -0.0017 & 0.0015 & 0.0066 \end{pmatrix}$$

We also find the variance-covariance matrix using simulations. To do that we generate 50 chains over time using the estimated parameters. The variance-covariance matrix using the simulations is given by:

$$\begin{pmatrix} 0.0087 & -0.0035 & -0.0054 & -0.0012 & 0.0047 & 0.0021 \\ -0.0035 & 0.0101 & -0.0058 & -0.0009 & 0.0026 & 0.0012 \\ -0.0054 & -0.0058 & 0.0194 & 0.0032 & -0.0086 & -0.0032 \\ -0.0012 & -0.0009 & 0.0032 & 0.0035 & -0.0011 & -0.0018 \\ 0.0047 & 0.0026 & -0.0086 & -0.0011 & 0.0089 & 0.0016 \\ 0.0021 & 0.0012 & -0.0032 & -0.0018 & 0.0016 & 0.0059 \end{pmatrix}$$

We see that the simulated variance-covariance matrix has close values to the partial likelihood, all entries having the same sign. We also look at the distribution of the estimators using the 50 samples. Figure 18 shows the parameter estimates approximately follow a normal distribution.

To estimate the desired probabilities, we generate samples (10000) from the parameter space using the mean of the parameters and variance-covariance matrix from a multivariate normal. To fix ideas suppose we want to compute the probability of no frost between (and including) the 274th day and the 280th day of the year. For every vector of parameters, we then compute the probability of observing $(0, 0, 0, 0, 0, 0, 0)$ exactly once given it was below zero on the 273th day and once it was above zero. In other words we compute

$$P(e(274) = 0, \cdots, e(281) = 0 | e(273) = 1),$$

and

$$P(e(274) = 0, \cdots, e(281) = 0 | e(273) = 0).$$

We also use the historical data to estimate $p_0 = P(e(273) = 1)$. Then the desired probability would be

$$\begin{aligned} P(e(274) = 0, \cdots, e(281) = 0) &= \\ p_0 P(e(274) = 0, \cdots, e(281) = 0 | e(273) = 1) &+ \\ (1 - p_0) P(e(274) = 0, \cdots, e(281) = 0 | e(273) = 0) & \end{aligned}$$

Then in order to get a 95% confidence intervals we use $(q(0.025), q(1-0.025))$, where q is the (left) quantile function of the vector of the probabilities.

Using the historical data, we obtain $p_0 = P(e(273) = 1) = 0.243$. Then for every parameter generated from the multivariate normal with mean and the above variance-covariance matrix we can estimate the two probabilities π_1 and π_2. We sample 10000 times from the multivariate normal, compute 10000 probabilities and take the 0.025th and 0.975th (left) quantiles to get the following confidence intervals for π_1 and π_2 respectively:

$$(0.28, 0.40),$$

and

$$(0.74, 0.85).$$

If we use the simulated variance-covariance matrix, we'll get the following confidence intervals for π_1 and π_2

$$(0.28, 0.40),$$

and

$$(0.75, 0.85),$$

which are very similar to the aforementioned intervals.

5 Possible applications of the models

To understand the potential applications of these models and results I contacted Dr. Nathaniel Newlands from AAFC (Agriculture and Agi-food Canada). He give the following insightful comments.

"Forecasted (probability of precipitation) is a leading indicator used by crop insurance companies. Probabilities of this kind (agroclimate) are typically most useful in early growing season by farmers in deciding planting dates and deciding on irrigation scheduling and ordering fertilizer and other kinds of inputs. Frost probability in latter growing season is critically important in deciding when to harvest crops before they have a higher potential for weather damage. So, essentially at the start and end of growing season, frost, precipitation (sometimes as a water stress index) and temp extremes are all informative for farmers and other decision makers in ag industry.

I would generally say that a broader set of probabilities like these are of special interest to the government side as they look for improving and/or developing new models, web portals and other tools to aid a wide array of the decision makers in the agricultural industry with their business decisions. Farmers (depending on what region of Canada they are in) are used to dealing with reoccurring weather and now climate change events, so often their viewpoint and decision needs are far more regionally specific than government which tries to balance regional with national needs and levels of risk to changing agroclimate.

The crop insurance industry is probably the most specific user of such information. For example, they base their insurance quotes for the event of precipitation on some specific times of the year."

Acknowledgement. We are indebted to L. A. Vincent from Environment Canada for providing the data on which the application in this paper is based. We also like to thank N. Newlands (AAFC) and R. Wright (AAFRD) for insightful comments. This work was partially supported by grants from the Natural Science and Engineering Research Council of Canada and the National Institute for Complex Data Structures.

References

[1] P. Embrechts, C. Klppelberg, and T. Mikosch. *Modelling extremal events for insurance and finance.* Springer, 2001.

[2] R. Hosseini. *Statistical Models for Agroclimate Risk Analysis.* PhD thesis, Department of Statistics, UBC, 2009.

[3] B. Kedem and K. Fokianos. *Regression Models for Time Series Analysis.* Wiley Series in Probability and Statistics, 2002.

[4] Hosseini R., Zidek J., and Le N. An analysis of alberta's climate. part ii: Homogenized data. Technical Report TR #246, Department of Statistics, UBC, 2009.

[5] Hosseini R., Zidek J., and Le N. r-th order categorical markov chains. Technical Report TR #248, Department of Statistics, UBC, 2009.