Dynamic Multivariate Continuous Data State-Space Estimation for Agrometeorological Event Anticipation

Philip Sallis, Sergio Hernández, and Subana Shanmuganathan

***Abstract*—This paper describes the selection of a state-space estimation method for application to the emerging research domain of agrometeorology. The work comes from a wider geocomputational research programme that relates to climate and environment monitoring and subsequent data analysis. In particular, the data currently being collected refers to meso-micro climates in vineyards across eight countries. It is terrestrial in kind, being in the context of near-ground truth continuous data. The time-related nature of the data, being continuous across a geo-spatial plane, gives rise to the need for mathematical models that are intrinsically spatio-temporal and while effective in their robust adequacy, are also computationally efficient. State-space models are considered a class of model within the time-series literature but they have some uniquely distinguishing features for continuous multivariate data representation. Ensemble Kalman Filter models are Bayesian based estimators of multiple realisations of state-spaces over time, so are proposed here as applicable to this analytical process domain.**

***Index Terms*—Geocomputation; estimation; agronomy; meteorology; sensors; monitoring telemetry.**

I. Introduction

The term *agrometeorology* has emerged to represent the area of research specifically related to climate studies in the agronomic domain. Agronomy itself can be considered as an over-arching term that combines research and practice relating agriculture, viticulture and horticulture. For each of these crop production contexts, timely and accurate environment impact information is critical for decision-making precision. Climate in particular, plays a significant role in determining crop yield and quality. This paper emanates from research in the field of *geocomputation*, which is concerned with techniques and technologies for both celestial and terrestrial environment monitoring and concomitantly with developing mathematical and statistical models using data collected from the monitoring to describe and illustrate likely event occurrences that in turn can be used for decision-making related to their anticipation. In particular, the work described here is in the context of a growing corpus of reported research using Bayesian methods for dynamic data models. A recently published comprehensive review paper relating to this research domain [1] describes spatial process modelling for univariate and multivariate dynamic spatial data, illustrating the use of a spatiotemporally varying coefficients and the enablement of predictive inference resulting from such an approach. They define a spatiotemporal framework as comprising two sets of equations; one for measurement and one for transition as new realisations of the model are generated. We will show here that dynamic models using means and covariances of estimated values from known data points may enable the generation of realisations of the model for every pixel point on a spatial plane between known truth value data point locations. We will demonstrate and interpolation method using a holistic approach where all means and covariances are retained for subsequent iterations of the model using *posterior* values as new *priors* in a recursive process that in effect is an example of implementing the transition equation proposed by these authors.

The wider research programme to which this paper relates provides a context for *agrometeorolgical* state-space event model selection. In essence it is an international collaborative research project relating in particular to vineyard management but with application to all aspects of agronomic research and practice. An architecture for a terrestrial wireless sensor network (WS) telemetry system has been designed, built and now deployed in some thirty locations across eight countries. Data collected in real-time from instruments in these locations provides continuous near-ground truth representations of climate and environment conditions. Details of the telemetry architecture, the sensor construction, calibration and deployment can be found at www.geo-informatics.org and in Ghobakhlou [2].

Processing the monitored data for a variety of purposes requires geostatistical analyses and mathematical modelling. Data is analysed using conventional non-parametric statistical methods but also by employing other geospatial interpolation methods such as inverse distance weighting and least squares based kriging and co-kriging. In order to incorporate time interval data into a probabilistic model or one for event anticipation (prediction) it is necessary to use an estimation procedure that provides greater precision than conventional time series methods, which are suitable for discrete data point value modelling but not for continuous stream data. Such an estimation procedure is the ensemble Kalman filter (EnKF) method. This method, which translates a recursive algorithm to a computationally intensive process, combines a set of equations for establishing *a priori* an n-dimensional matrix that on every iteration of the model as new information is received concerning data point values over time, generates a fresh

Manuscript received June 30, 2012; revised August 30, 2012.

P Sallis is with Geoinformatics Research Centre (GRC), School of Computing and Mathematical Sciences, Auckland University of Technology (AUT), Auckland, New Zealand (e-mail: philip.sallis@aut.ac.nz).

S. Hernández is with Laboratorio de Procesamiento, de Información Geoespacial,Universidad Católica del Maule, Talca, Chile. (email:shernandez@ucm.cl).

S Shanmuganthan is with Geoinformatics Research Centre, School of Computing and Mathematical Sciences, Auckland University of Technology, Auckland, New Zealand. (e-mail: subana.shanmuganthan@aut.ac.nz).

realisation of the model, such that *posterior* values become the new *priors* for each instance of the model generation. Because this method enables us to produce a new realisation of the model on every time-base iteration and because these realisations depict estimations for each pixel point between known location data points (the locations where near-ground truth is known from data collected by sensors at those locations) we can generate robust spatio-temporal estimation models. Some early work with this essentially Bayesian based estimation technique can be seen at Sallis and Hernández [3, 4].

II. The Wireless Sensor Network

The Wireless Sensor Network (WSN) designed and built for this research has been focused on collecting, logging and processing near-ground truth data. From this data some further pre-analytical values are derived and in the post data acquisition phase others are computed. See for example, Shanmuganathan et al [5]. The configuration used for implementing the WSN architecture can be seen in Ghobakhlou [6].

III. Ease Analytical Methods

Agrometeorological models have their genesis in weather forecasting. Day-to-day management and strategic planning for crop production development have used these models successfully in the past [7, 8] despite the natural world reality that random events such as floods and droughts mitigate against some precision in these models [9].

Distance into the future (range) is a major contributing factor for precision as is resolution; the latter being most important for meso and micro climate forecasting due to the great constriction on weather variability in spaces with lower radial values than those where the state-space is larger [10]. This means that more physical modelling data is required for determining the state of the atmospheric values, which are more prone to rapid change than plant or soil variable values. Forecasting non-stationary dynamic signals has been the preoccupation of scientists since the early 1960's but now with continuous data recorded and transmitted more readily with contemporary data communications technologies, the temporal aspect of such sampling has encouraged researchers into considering new approaches. This is most obvious when we observe that the focus in meteorology using data assimilation was mainly based on deterministic non-linear filters, where the dynamic model is a perfect representation of the physical system [11]. A significant issue here though, is that in weather forecasting systems based on data assimilation, their sensitivity to initial conditions puts fundamental limits on the prediction potential of available models, which means that for any interpolation where changes occur over time, we need to consider how the changes affect one data point with another.

In this regard, we consider that change point analysis [12] could improve the detection of variable value shifts, especially over large historical sets of data. We also consider this precision to be essential for micro-climate modelling where time intervals are typically small because of the short topographic distances between data points and yet condition changes can be large. We propose that in such situations when we model large historical climate data sets we need to ask the questions, did a change occur? Did more than one change occur? When did the changes occur? With what confidence can we state that the changes did occur? So we have adopted this analytical framework approach mindful that change point analysis is a method capable of detecting multiple changes and for climate variation plotting we need to incorporate multiple levels of abstraction from the data we are observing and use this as a baseline approach for interpolation purposes.

Linear least squares estimation algorithms in the form of the geo-statistical Kriging methods [13] are probably the most popular for geospatial interpolation because they enable the prediction of unknown data point conditions (values) to be determined from a known set of values from neighbouring data points. In this way we can model changes across a plane with a high degree of value expectation certainty. Data points with geo-referenced values for latitude and longitude (x and y) can be interpolated with their elevation data (z) to provide terrain maps in three dimensions. The greater the number of data points the better the expectation confidence. So when values from three or more data points are known, greater precision can be observed in the output value.

Taking a holistic approach in this way means we are less likely to lose valuable descriptive data for later interpretation. The *Ensemble* approach [14] for modelling geospatial data is such a holistic method. Indeed, it is an intrinsic characteristic of the approach. These methods provide even greater estimation precision to the data model because they utilise a multiple analysis approach and apply several hypothesis algorithms to a single learning proposition. A refinement of the hypotheses is possible when taking this approach because the intrinsic learning algorithm of the method prunes so-called *weak learners* to focus on the strength of results produced by one of them. This method is more computationally intensive than using for example, a single supervised neural network algorithm but its precision appears to be superior.

While modelling of historical climate data is useful for anticipating future trends, there is a further challenge for modelling continuous rather than discrete point data. The temporal aspect of these geospatial models adds considerably more complexity to the processing, which reflects the inherent complexity of the data brought on by random events in Nature. We are continuing to explore ways to approach and address this challenge.

Researchers in Chile [15] have used mesoscale models such as MM5 and MOS in the Central Valley wine growing region to forecast surface meteorological and agro-climatic variables. The authors of this research reported a spatio-temporal interpolation method for temperature, wind speed, relative humidity, and daily solar radiation in grid cells with a spatial resolution of up to fifteen (16) kilometres.

In other examples, synoptic and planetary circulation models [17] have provided a large scale hydrodynamic approximation to the climatic patterns while mesoscale circulation models are used to characterize horizontal scale, which are smaller than the synoptic scale.

In yet another context, since the 1960's the signal

processing community has been interested in stochastic linear filters for signal tracking with uncertain observations. In this case, the dynamic model is not perfect and it is considered as being corrupted with random noise. More recently and as reported in Sallis and Hernández [3] the Ensemble Kalman filter (EnKF) [18] has been proposed in data assimilation situations to model uncertain initial conditions in numerical weather prediction. The EnKF overrides the linearity assumption of the standard Kalman filter by using a Monte Carlo approximation of the optimal probability forecast. Because of the inherent so-called 'curse of dimensionality problem' of stochastic approximation methods such as with a sequential Monte Carlo, the EnKF uses a low-rank approximation to the covariance of the posterior density, which also introduces spurious correlations in the filter estimates. A Digital Elevation Model (DEM) and the distance from the sea data are used in the temperature interpolation [19]. The interpolated values will then be used as observations for a sequential Monte Carlo method with an added layer of sophistication coming from an imbedded parameter selection algorithm for estimating the dynamic climate pattern of the state-space we are observing.

It should be noted here that obtaining a large set of complete data including elevations is not as simple as it might seem. In our ongoing work we are assembling similar data sets from both Chile, see [20] for an example of integrating macro and micro climate data for frost prediction in Chilean vineyards and New Zealand for comparative purposes to test the interpolation method and observe the output models for similarities and differences. The accession of this data is expected to provide a large volume of continuous data suitable for our ongoing work with change point analysis and ensemble methods.

IV. Units

The results from these experiments were previously published in Sallis and Hernández [3], which are repeated here for completeness of both method description and evidential purposes. We wanted to test our concept using the Ensemble Kalman Filter approach, so we examined some data available to us from Croatia. This country, situated between the Mediterranean Sea and the high mountains on its border with Hungary, provides an interesting context for examining climate variability in a relatively small general location. In these experiments we confined our study to temperature variation and observed changes over time across the data plane. We had data from 123 weather stations located throughout the country with data for near-ground truth from these stations for a one year period. We employed the EnKF method to estimate values for intermediate unknown data points based on interpolations using the known data points. We used daily surface temperature data from these stations. The data includes location, elevation (average mean sea level), distance-from-the-sea (as the crow flies), surface temperature and the sampling time.

Near ground truth data for each of the 123 weather stations over a one year period provided us with sufficient data to test the EnKF method. We set out to test the method *a priori* and in order to investigate its veracity. The computational resources required for software development and execution of the model are substantial. Therefore, we did this only for a single data point over the one year period in order to observe the dynamics of the algorithm and the ensuing model. For this study we interpolated the data using a polynomial regression method in order to minimise overt variance values in the processing. The interpolated values were then used as initial observations for input to an EnFK model.

The polynomial regression results using a probability density function are illustrated in Fig. 1 below. They indicate a high proportion of errors outside of the Gaussian distribution curve.

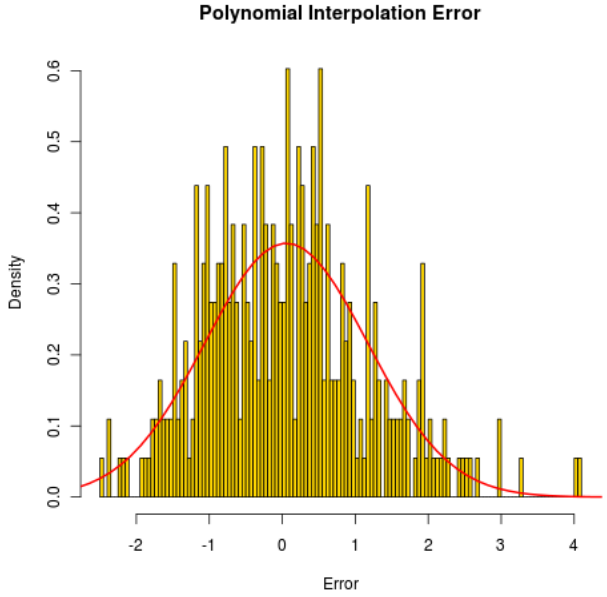

Fig. 1. Fitting a Gaussian curve to the interpolated data

This Gaussian curve-fitting exercise illustrates an error distribution for a single sensor over one year. The probability density function applied here indicates a maximum from zero value of 0.6 whereas within the normal curve the maximum is 0.34, nearly one half less. A goodness-of-fit (Shapiro Test) shows a poor fit of this data with a *p* value of 0.0007721.

V. State-Space Data Modelling

State-space modelling for land surface temperature forecasting [16] is an integral component of our approach. A state-space model contains two equations for describing the dynamic behaviour of the system and the observational process. As seen below the state-space representation is conceptually a graph for sequential probabilistic inference over a partially observed stochastic process. The state **x** is an unobserved first-order Markov process and the observations are conditionally independent given the state process.

$$x_k = f(x_{k-1}, v_k) \quad \text{process equation}$$
$$z_k = g(x_k, w_k) \quad \text{observation equation}$$

The state of the system x_k at time k is a Markov process observed via the measurement z_k. The noise sources v_k and w_k are assumed as being mutually independent and identically-distributed (i.i.d.) sequences of random variables,

which are also independent of the state and the observations x_k and z_k respectively. The functions f and g represent possibly non-linear mappings from x_{k-1} to x_k and from x_k to z_k respectively.

When the state-space is linear with Gaussian additive noise, the well-known Kalman filter achieves the solution for the optimal estimation problem. The Kalman filter is the most popular technique for handling linear models with Gaussian distributed noise. When the state-space can be written as a linear dynamic model with zero-mean Gaussian noise sources v_k, $N(0,Q_k)$ and w_k, $N(0,R_k)$, the posterior density is also Gaussian so it can be completely parameterized by its mean and covariance. Let A_k and B_k be two matrices defining a linear transformation for the process and observation equations. Q_k and R_k represent the process and observation noise covariance respectively. The linear Gaussian state-space with a seasonal component can be written thus,

$$x_k = A_k x_{k-1} + \frac{2\pi}{T} C_k + v_k$$

$$z_k = B_K x_k + w_k$$

The Kalman filter computes the optimal conditional mean and covariance of x_k by recursively predicting and updating a Gaussian belief distribution. The recursive method is optimal since using the following equations minimize the mean square error of the observations and the predicted state. The term S_k denotes the covariance of an innovation matrix $\varepsilon_k = z_k - B_k x_{k|k-1}$ that generates a sequence of uncorrelated terms. The superscript T denotes matrix transposition and K_k is the so-called Kalman gain. Both terms S_k and K_k can also be written as,

$$S_k = B_k \sum_{k\ |k-1} B_k^T + R_k$$

$$K_k = \sum_{k\ |k-1} B_k^T S_k^{-1}$$

When applied to the sample data being used here, the EnKF can be seen to perform well in terms of producing a robust state space estimation model. The EnKF model was tested against reported near ground 'truth' using estimates (predictions) of Mean Day Temperature (*MDTemp)* on each of 365 days for a single data point and produced the results illustrated in Fig. 2. These results were output from 1000 realisations of the ensemble.

We observed that the final ensemble (Blue in the second or lower graph) was practically identical to the near ground truth (Red line in the graph above) and that although the interpolated values (results from the polynomial regression shown as Red Circles in the second or lower graph) were dispersed across the ensembles, the 1000 realisations (Grey in the second or lower graph) followed a pattern of distribution that clearly indicates the spectrum of estimates made consisting of their individual means and variances. The final realization of the model is clearly similar to the near ground truth in the first or upper graph and that is the significant result reported here because it indicates a potential robustness of method for which, the EnKF appears responsible. We propose this to be a satisfactory result.

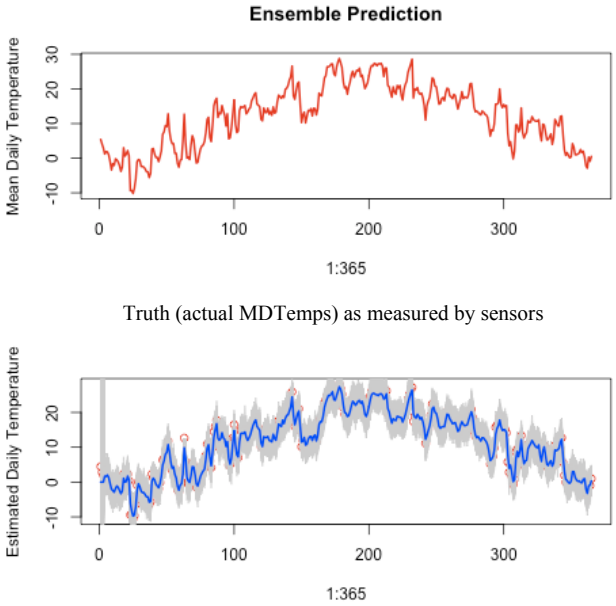

Fig. 2. EnKF model compared with truth

VI. Conclusion

The Ensemble Kalman Filter (EnKF) method has been chosen here to deal with the spatio-temporal estimation problem. Confidently identifying and determining values for discrete data points across a three dimensional plane to model climate variation is a non-trivial challenge for any single interpolation method. Outlying values that may not conform to the expected variations to a mean may in fact, be significant indicators of a change point yet to be observed. Kriging for instance, would prune such a value and complete the interpolation without including it in the cluster of predictors for new data point instances. As our previous work demonstrates, ensemble methods provide a multi algorithmic approach that does not discard any values until computations of all possible permutations of the data are exhausted. They also allow for a temporal variable to be meaningfully incorporated into the model without distorting the intrinsic geospatial properties of the former interpolation methods.

Although it is too early for us to state emphatically (or empirically) that the EnKF approach is entirely reliable but it appears to predict accurately against known truth. Using a Kalman Filter to maintain data integrity and reduce noise in the data set during computation produces a clean and reliable model and a result.

Our current work is concerned with parameter selection for model algorithm refinement with which we hope in the future to illustrate by comparing the two model outputs when sufficient appropriate data is available from the WSN referred to here in the context of spatial state space estimating.

REFERENCES

[1] Gelfand, A E, Banerjee, S and Gamerman, D. (2005), "Spatial Process Modelling for univariate and multivariate dynamic spatial data," Environmetrics 2005, vol.16, pp.465-479

[2] Ghobakhlou, A., Sallis, P., Diegel, O., Zandi, S. and Perera, A. (2009). "Wireless sensor networks for environmental data monitoring" IEEE Sensor 2009 Conference 25-28 Oct 09, Christchurch, NZ.

[3] Sallis, P. and Hernández, S. (2010) "Ensemble interpolation methods for spatio-temporal data modelling". In *Proceedings of the 4th European Symposium on Mathematical Modelling and Simulation*, Pisa Italy, 17-19 Nov 2010(in print).

[4] Sallis, P. and Hernández, S. (2011) "Geospatial state space estimation using an Ensemble Kalman Filter," *International Journal of Simulation, Systems, Science and Technology*. (in print)

[5] Shanmuganathan, S. Sallis, P. and Narayanan, A. (2009) "Unsupervised Artificial Neural Nets for Modelling the effects of climate change on New Zealand Grape Wines," *Proceedings of 18th World IMACS/MODSIM Congress*, Cairns, Australia, 13-17 July 2009, pp: 803-809

[6] Ghobakhlou, A., Perera, A., Sallis, P, and Zandi, S. (2009), " Modular Sensors for Environmental Dat Modelling".

[7] Caprio, J M, and Quamme H A. (1998), "Weather conditions associated with apple production in the Okanagan Valley of British Columbia," Agriculture and Agri-Food Canada Pacific Agri-Food Research Centre, Summerland, British Columbia, Canada, V0H 1Z0 1998, vol. Contribution no.1075 129-137.

[8] Van Leeuwen, C., Friant, P., Choné, X., Tregoat,O., Koundouras,S., and Dubourdieu,D. (2004), "Influence of Climate, Soil, and Cultivar on Terroir," Am. J. Enol. Vitic. 2004, pp.55:3:207.

[9] Delyam, A.M., "Chaotic Climate Dynamics," Lunivar Press, 2007. ISBN-13 978-1-905986-07-1.

[10] Jones, G V, and Davis, R E,. (2000), "Climate Influences on Grapevine Phenology, Grape Composition, and Wine Production and Quality for Bordeaux, France," Vols. Am. J. Enol. Vitic., vol. 51, no. 3, 2000 pp249-261.

[11] Holton,J. "An introduction to dynamic meteorology". *Academic Press*, 2004.

[12] Berger, J. O., De Oliveira, V. and Sansó, B. (2001). "Objective Bayesian analysis of spatially correlated data," Journal of the American Statistical Association, 96, pp: 1361—1374.

[13] Drignei, D. "A kriging approach to the analysis of climate model experiments," (2009) *Journal of Agricultural, Biological and Environmental Sciences*. Springer New York 2009 vol. 14(1) pp 99-114. ISBN 1085-7117 (Print) 1537-2693 (online).

[14] Okun, Oleg; Valentini, Giorgio (Eds.) "Supervised and Unsupervised ensemble methods and their applications" in *Springer series*, Studies in Computational Intelligence, vol. 126 2008

[15] Barry, RJ and Carleton, AM. "Synoptic and dynamic climatology," Routledge, 2001.

[16] Silva, D, Meza, FJ and Varas E. "Use of mesoscale model MM5 forecasts as proxies for surface meterorological and agroclimate variables," Cienc. Inv. Agr. [online], 2009, vol. 36, no.3.

[17] Grewal, M. S. "*Kalman Filtering: Theory & Practice,"* Englewood Cliffs, NJ: Prentice-Hall, 1993.

[18] Petrosyan, AS. GIS in meteorology and climatology. The needs and the challenges. European Geophysical Society. XXVI General Assembly. Nice, 25-30 March 2001.

[19] Sallis, P., Jarur, M., and Trujillo, M.(2009). "Frost prediction characteristics and classification using computational neural networks," in *Australian Journal of Intelligent Information Processing Systems (AJIIPS)* vol. 10.1, 2008 (ISSN 1321-2133) pp50-58.

[20] National Institute of Water & Atmospheric Research. The National Climate Database, National Institute of Water & Atmospheric Research,.[Online]. Available: http://cliflo.niwa.co.nz/