Geospatial state space estimation using an Ensemble Kalman Filter

Philip Sallis
Geoinformatics Research Centre
School of Computing and Mathematical Sciences
Auckland University of Technology
Auckland, New Zealand
psallis@aut.ac.nz

Sergio Hernandez
Laboratorio de procesamiento
de información geoespacial
Universidad Católica del Maule
Talca, Chile
shernandez@ucm.cl

***Abstract*—Incorporating temporal (continuous) data into more common discrete data point geospatial models is necessary for dynamic real time model building. The models are otherwise limited in their use for numerical modelling, simulation and the prediction of climatic states over time. By adopting a Bayesian approach it is shown here to be possible to estimate the dynamic behaviour of unobserved climate patterns forward over time using state-space representations. The recursive state-space Ensemble Kalman Filter (EnKF) is proposed here as a solution to the spatio-temporal modelling problem. This or a more general sequential Monte Carlo method could be used for the estimation procedure but the Ensemble Kalman Filter is observed as producing robust models for the temporal geospatial domain. The EnKF approach is outlined here, with some sample data analysis to illustrate its application and a description of a real-time climate modelling telemetry architecture for data acquisition and model provisioning.**

Keywords-climate modelling, interpolation, ensemble methods, kalman filter, GIS, Wireless Sensor Networks

I. INTRODUCTION

Gazing at the sky in an attempt to ascertain upcoming weather conditions was probably practiced by Neolithic man for decision making related to hunting, travelling or other activity and has been for a variety of purposes ever since. Modern meteorology owes much to the Danish physicist Hans Christian Orstead (1777-1851) who discovered electromagnetism and since then to many others who have refined the science of weather forecasting. Today as much as any time in the past, weather forecasting is a highly influential tool in decision making for agriculture, horticulture and viticulture. It is this in this context that the work outlined here was undertaken. A wider programme of research relating to environment monitoring and modeling that incorporates the data modeling issues considered by this paper will be described. In particular, the paper refers to the design, construction and implementation of a real time data communications architecture that incorporates a wireless sensor network telemetry of terrestrial sensors for use in agronomic meso-micro climate monitoring situations. Contemporary communications technologies that can provide real time near-ground truth data acquired from both celestial and terrestrial sensors as input to dynamic computational models can produce increasingly accurate weather forecasts. For macro, meso and micro climate models, Geographic Information Systems (GIS) can be built to provide information about topographic data such as elevation and distance to oceans or water reservoirs.

Increasingly available globally, this data has begun to provide new methods and applications with greater potential for robust modelling. By using geographic information together with terrestrial measurements from weather stations, the spatial and temporal scales of the climatic variables can be analyzed by interpolation and forecasting methods. Meteorological models for weather forecasting have been proved to be useful for strategic planning and management in agriculture [1, 2] even though some precision is mitigated by random events in nature [3].

The success of agro-climatic models is dependent on the range and resolution of the forecast [4] so adequate physical modelling is required for determining the state of the atmospheric values. In the early 1960's, researchers and engineers from different fields were concerned with the problem of forecasting non-stationary dynamic signals. 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 [5] Nevertheless, one of the main issues in weather forecasting systems based on data assimilation is their sensitivity to initial conditions, which puts fundamental limits on the prediction.

In general terms interpolation methods for geospatial data modelling are best appreciated when we consider how changes occur from one data point to another over time. Change point analysis [6] improves the detection of variable value shifts, especially over large historical sets of data. We 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. For this reason, when we model a large historical climate data set (described below) we 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? We have adopted this analytical framework approach because 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.

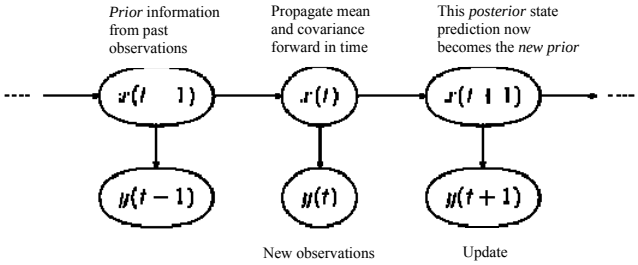
II. INTERPOLATION METHODS

Linear least squares estimation algorithms in the form of the GeoStatistical Kriging methods [7] are probably the most popular for geospatial interpolation because they enable the prediction of unknown data point conditions (values) determined from a known set of values from neighboring data points. This way we can model change 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 a terrain map 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.

When we describe *Ensemble* methods [8] for modelling geospatial data, we do so holistically, which is an intrinsic attribute 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.

While modelling of historical climate data is useful for anticipating future trends, there does exist 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, which after all, is a facet of Nature where complexity also includes random events, leading stochastic models on catastrophic occasions to reflect the properties of chaotic manifestations. This challenge forms part of our ongoing work. Synoptic and planetary circulation models [9] provide 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. Because of the lack of a high resolution meteorological network, mesoscale models such as MM5 and MOS [10] have been used for forecasting surface meteorological and agroclimatic variables in central Chile. 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 (15) kilometers.

In another context, 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. The Ensemble Kalman filter (EnKF)[11] 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. The recursive sequencing nature of the method can be depicted using Markov notation thus:



This is a three step recursive process such that **T** is modified iteratively by new information ⟶ T|T-1⟶T+1|T⟶T+1|T+1⟶ where an *a priori* state **T** given **T-1** is updated with a mean and covariance from new observations (realisations) at **T+1** (given **T**), which then as **T+1** becomes the forecast estimate (*posterior prediction)* and is returned as the *prior* for the next iteration of the model, when **T+1** becomes **T**.

III. APPLICATION OF THE ENSEMBLE METHOD

Experiments with forecasting climatic states using spatial interpolation and ensemble methods is ongoing in the research programme to which this paper refers. Because it is an early stage project only preliminary results are available (as outlined below) but we offer our approach as being conceptually appropriate for the problem domain. Only now is the wireless sensor network (WSN) intended

for model propagation being installed and a large set of data will eventually be available for analysis and modeling. We intend to use a Digital Elevation Model (DEM) and the distance from the sea generated by a Geographic Information System (GIS) to interpolate temperature from the weather stations [12]. The interpolated values will then be used as observations for a sequential Monte Carlo method for estimating the dynamic climate pattern.

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 [13] for an example of integrating macro and micro climate data for frost prediction in Chilean vineyards and New Zealand [14] for comparative purpose to test the interpolation method and observe the output models for similarities and differences. A wider research programme to which the work described in this paper relates, consists of a wireless sensor network (WSN) with 28 data collection stations in 8 countries. Each station has a minimum of three sensor arrays, each consisting of 17 sensors for atmosphere, climate, plant and soil. This research relates to monitoring and modeling in viticulture, horticulture and agriculture for precision information provision with a view to effective sustainable management [15]. 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.

In order to test our concept using the Ensemble Kalman Filter approach, 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 say temperature change over time across the data plane. We wanted to observe how the EnKF method performed against the known data from 123 weather stations, when the values for intermediate points were estimated. 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. A map of Croatia illustrates its particular geographic location as in Fig. 1

Figure 1. A topographical map of Croatia

In Fig. 2 a digital elevation model (DEM) image of Croatia illustrates the location of the weather stations from where the data in this paper were acquired. This depiction shows the potential interpolation space for our study, which consists of all pixel locations across the geospatial plane where near ground temperature truth exists for known data points.

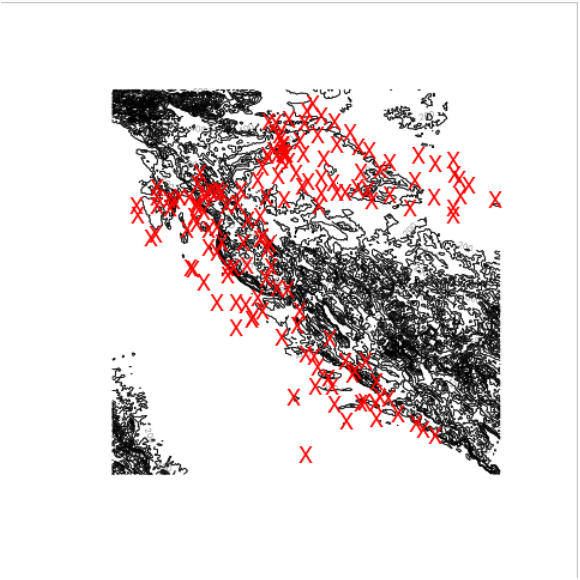


Figure 2. A digital elevation model for Croatia with superimposed weather station locations

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 dymanics 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. 3 below. They indicate a high proportion of errors laying outside of the Gaussian distribution curve.

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.

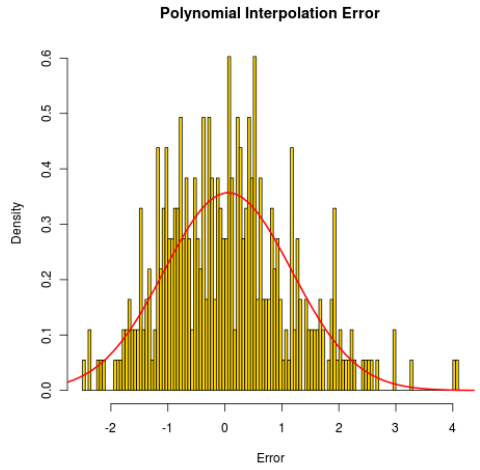


Figure 3. Fitting a Gaussian curve to the interpolated data

III. SPATIAL INTERPOLATION OF CLIMATIC VARIABLES

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,

$$S_k = B_k \sum_{kk-1} B_k^T + R_k$$
$$K_k = \sum_{kk-1} B_k^T S_k^{-1}$$

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. 4. 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 and a useful addition to the nascent discussion of this modeling approach in the geospatial interpolation research community.

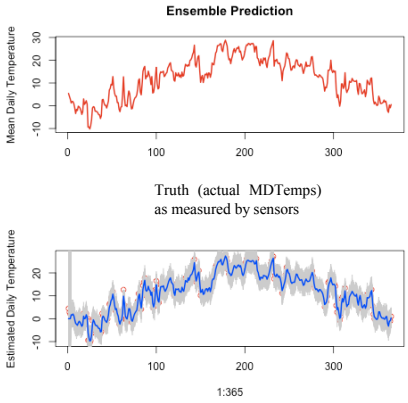


Figure 4. EnKF model compared with known near ground truth

V. CONCLUSIONS

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. 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. 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 so we can expect it will provide us with robust estimated values for unknown states, such that we can have some confidence that the information we produce is reliable. 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. We hope in the future to illustrate this by comparing the two model outputs when sufficient appropriate data is available from the WSN referred to here. Our work continues with installation of the WSN and ongoing exploration of innovative interpolation methods in order to gain insights to their dynamics and usefulness for robust model building in the context of spatial state space estimating.

REFERENCES

[1] 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

[2] 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

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

[4] 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.

[5] Holton,J. An introduction to dynamic meteorology. Academic Press, 2004.

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

[7] 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)

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

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

[10] 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

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

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

[13] 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) volume 10.1, 2008 (ISSN 1321-2133) pp50-58. Also published in M. Kppen et al. (Eds.): ICONIP 2008, Part I, LNCS 5506, 2009. Springer-Verlag Berlin Heidelberg 2009. pp. 1211-1220.

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

[15] 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, New Zealand.