Xjenza Online - Journal of Malta Chamber of Scientists
http://www.mcs.org.mt/
Doi: http://dx.medra.org/10.7423/XJENZA.2013.1.06

Student Research Article

COMPUTATIONALLY EFFICIENT ESTIMATION OF HIGH-DIMENSION AUTOREGRESSIVE MODELS - WITH APPLICATION TO AIR POLLUTION IN MALTA

Luana Chetcuti Zammit[1], Kenneth Scerri[1], Maria Attard[2] and Thérèse Bajada[2]
[1]Department of System and Control Engineering, University of Malta, Msida, Malta
[2]Institute for Sustainable Development, University of Malta, Msida, Malta

Abstract. The modelling and analysis of spatio-temporal behaviour is receiving wide-spread attention due to its applicability to various scientific fields such as the mapping of the electrical activity in the human brain, the spatial spread of pandemics and the diffusion of hazardous pollutants. Nevertheless, due to the complexity of the dynamics describing these systems and the vast datasets of the measurements involved, efficient computational methods are required to obtain representative mathematical descriptions of such behaviour. In this work, a computationally efficient method for the estimation of heterogeneous spatio-temporal autoregressive models is proposed and tested on a dataset of air pollutants measured over the Maltese islands. Results will highlight the computation advantages of the proposed methodology and the accuracy of the predictions obtained through the estimated model.

Keywords Data-driven modelling; Spatio-temporal autoregressive (STAR) models; Sparse datasets

1 Introduction

Mathematical modelling and analysis is an indispensable tool in the study of both natural and anthropogenic phenomena ranging from the spread of social media to the analysis of the human brain. Due to complexity of the interactions involved, mathematical modelling through the use of known physical, biological, chemical or economic laws is often unfeasible. Nevertheless, such systems often provide large datasets of measurements that indirectly describe the relationships involved. Thus, in a data-driven modelling framework, models are extracted directly from the data through a process of successive estimation and validation until a required level of predictive accuracy is obtained. Such strategies have proved useful in various applications including biology (Shen et al. 2006, Shen et al. 2008); ecology (Ikegami and Kaneko, 1992; Nikolus and Gonzalez, 2002); meteorology (Amani and Lebel, 1997; Berliner et al. 2000); physics (Guo et al. 2006; Kessler et al. 1990); econometrics (Cliff et al. 1974; Giacinto et al. 2006) and chemistry (Shibata et al. 2002; Reiter, 2005).

One of the most widely accepted mathematical descriptions for data-driven modelling is the family of time-series models (Cliff et al. 1974; Martin et al. 1975). In their simplest form, Auto-Regressive (AR) models aim to capture the temporal relationships between successive measurements allowing for the description to consider data as far back in time as deemed fit for each application. Both stochastic and deterministic variables contributing to the measurements can be included through the use of Auto-Regressive Moving-Average (ARMA) or Auto-Regressive with eXogenous input (ARX) models, respectively. Although most applications consider only the linear relationships among the dataset, nonlinear variants such as the Nonlinear AR (NAR), Nonlinear ARMA (NARMA) and the Nonlinear

Correspondence to: L. Chetcuti Zammit (lche0003@um.edu.mt)
Received: 22/1/2013 - Revised: 5/3/2013 - Accepted: 10/3/2013 - Published: 31/03/2013

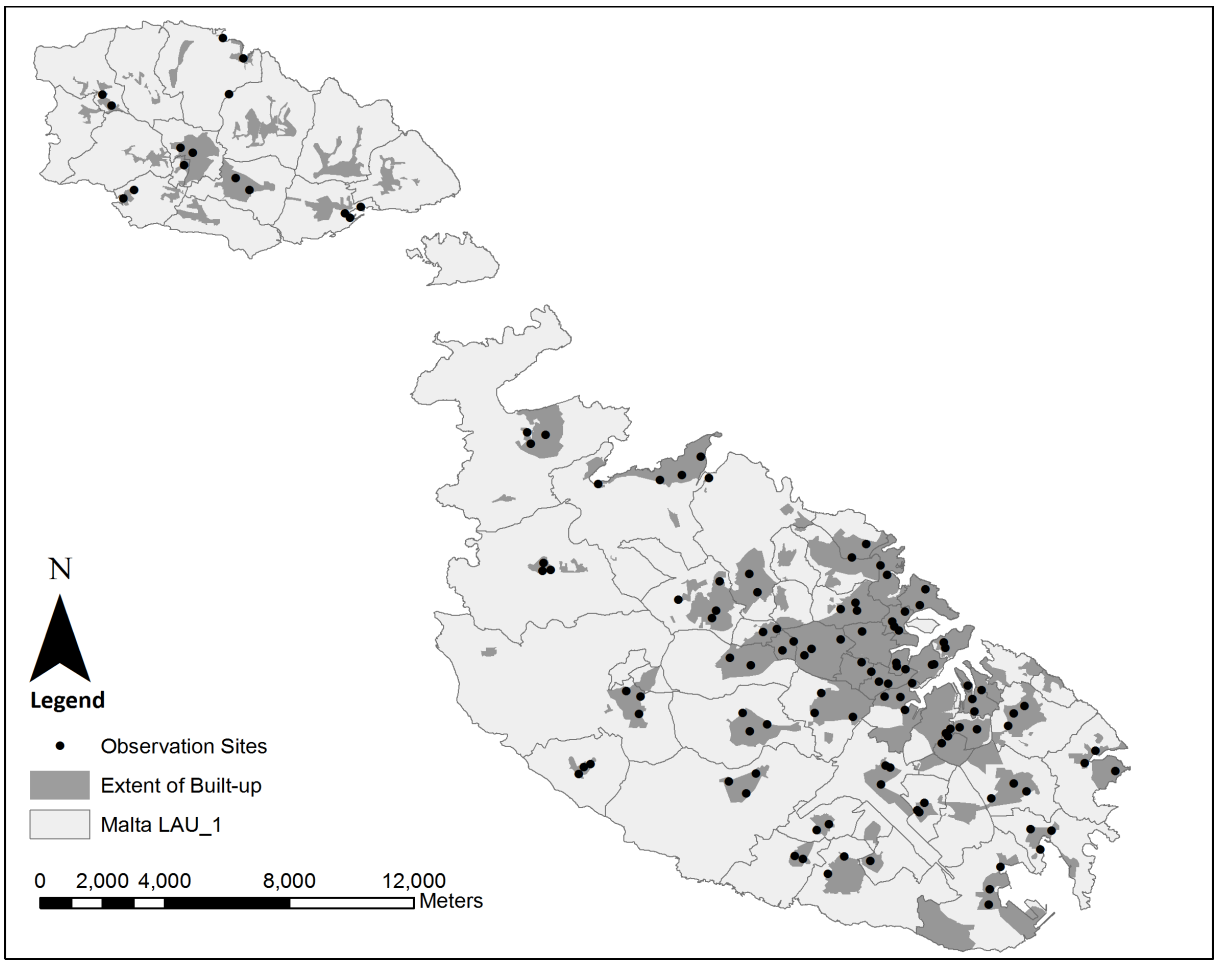

Figure 1: The Distribution of Passive Diffusion Tubes in Malta.

ARX (NARX) models have been proposed (Leontaritis et al. 1985a, Leontaritis et al. 1985*b*).

Multivariate AR (MAR) models are also widely used in applications where multiple measurements related to the same behaviour are being gathered. Such models are widely applicable to the emerging field of spatio-temporal modelling and analysis (Cressie et al. 2011), where data related to the same behaviour is gathered from various spatial locations. In such applications, modelling for both analysis and prediction will surely benefit from the inclusion of spatial interactions together with temporal dynamics. This observation contributed to the development of the Spatio-Temporal AR (STAR) models first proposed in (Pfeifer et al. 1980a, Pfeifer et al. 1980b) as one of the first tools to capture spatio-temporal behaviour from data.

Any data-driven modelling procedure follows a three step-strategy; starting with pre-analysis, filtering and model structure choices, prior to estimation of the model parameters and finally validation of the results obtained. Each step is well documented in literature for both temporal (Ljung et al. 1999; Chatfield et al. 2004) and spatial processes (Cressie et al. 1993) separately, but far less literature is available on methods for spatio-temporal modelling. This can be mostly attributed to the vast datasets usually associated with spatio-temporal studies resulting in significant computational challenges in the data-driven modelling procedure. Most significantly, methods widely used for estimation such as the least square criterion, require the repeated evaluation of a matrix inversion of the dimension of the spatial domain (Lutkepohl et al. 2005; Peña et al. 2001) which is intractable in higher-dimensional problems.

Thus in this work we present a method first proposed in (de Luna et al. 2005) in a MVAR setting for the efficient estimation of AR model parameters for high-dimensional problems. This method was set in a spatio-temporal setting in (Chetcuti Zammit et al. 2011) where, the spatial dependency is usually significant only among measurement sites located in close vicinity. Here this work is generalized to consider any STAR model and tested on a dataset of air pollution measurements taken over the Maltese Islands. Due to the sparsity of the multidimensional parameters in such applications, the

proposed method simultaneously estimates the model parameters together with the number of parameters required to capture the significant spatial relationships in the data. This reduction in the number of model parameters to be estimated allows for improvements not only in the computational demands but also in the parameter accuracy.

The remainder of this paper is structured as follows. First, a theoretical overview of spatio-temporal autoregressive models is given, followed by the proposed methodology. This methodology is validated on a dataset of air pollutant concentrations gather in Malta over the period 2004 to 2010. Some remarks on the spatio-temporal behaviour of these pollutants are then presented and finally conclusions are drawn on the applicability of the proposed methodology to such applications. Finally, various other possible future additions are identified and briefly discussed.

2 Spatio-Temporal Auto-Regressive (STAR) Models

A STAR model of order (p, q) is given by:

$$\begin{aligned}\mathbf{z}(\mathbf{s},t) = \mathbf{A}_1\mathbf{z}(\mathbf{s},t-1) + \mathbf{A}_2\mathbf{z}(\mathbf{s},t-2) + K \\ + \mathbf{A}_p\mathbf{z}(\mathbf{s},t-p) + \mathbf{e}_t \end{aligned} \quad (1)$$

where, $\mathbf{z}(\mathbf{s},\mathrm{t}) \in \Re$ represents the spatio-temporal process of interest as a stationary temporal series (Chatfield 2004), p denotes the temporal order, q denotes the full spatial order, $\mathrm{A_i} \in \Re^{\mathrm{q}\times\mathrm{q}}$ are the autoregressive parameters, and $\mathbf{e}_\mathrm{t}$ denotes white noise with the expectations $E[\mathbf{e}_\mathrm{t}] = \mathbf{0}$, $E[\mathbf{e}_\mathrm{t}\mathbf{e}'_\mathrm{t}] = \Sigma$ and $E[\mathbf{e}_\mathrm{t}\mathbf{e}'_\mathrm{t}] = \mathbf{0}$ for $u \neq t$.

Classical methods for the estimation of the model (1), require the inference from data of the temporal order p and the $(q^2 \times p)$ model parameters $\{\mathbf{A}_\mathrm{i}, \mathrm{i} = 1, \ldots \mathrm{p}\}$. In a frequentist statistical setting, the model parameters are usually estimated by the maximum likelihood or the least squares criteria. For the linear model (1), it is well known that both these criteria give equal estimates (Ljung, 1999). Nevertheless, both methods suffer severely from the curse of dimensionality, with the number of scalar parameters increasing quadratically with the number of spatial locations.

An adequate temporal model order is usually identified by the use of various model selection criteria such as the Akaike Information Criterion (AIC) (Akaike 1974) or the Bayesian Information Criterion (BIC) also called the Schwarz Criterion (Schwarz 1978). Both these criteria aim to identify the temporal order that best satisfy the principle of parsimony (Chatfield 2004), that is, the temporal order which best balances the model demands for generality and prediction accuracy. Nevertheless, this modelling strategy requires continuous user intervention with the user ultimately deciding on the preferred model order after fitting models of various dimensions.

3 The Proposed Methodology

The methodology being proposed in this work aims to mitigate the two limitations highlighted above by: limiting the number of model parameters to be estimated based on the known independence of non-neighbouring measurements and provide a single algorithm to identify both the models order and the system's parameters without any user intervention. This method was first proposed in (de Luna et al. 2005) and is here being modified, to make use of the natural spatial ordering of measurements based on their vicinity as highlighted in Algorithm 1.

Notes:

1. The spatial nature of the spatio-temporal phenomena being considered provide a natural ordering for the sites based on their spatial vicinity. Although such an interpretation is advantageous in various applications with diffusive behaviour (such as the pollution application being considered in this study), it does not allow for the identification of long distance interactions present in some biological applications such as the spatio-temporal modelling of the electrical activity inside the human brain.
2. The comparative measures used in this work are the AIC and BIC, although the methodology can easily accommodate any other measure (such as the modified AIC (McCullagh et al. 1989) or partial correlation measures (Peng et al. 2009)); as deemed appropriate for the particular application.
3. The user is only required to the make choices for the maximum temporal and spatial orders to be considered, that is pmax and qmax, respectively. If the final model choice is given by these maximum values, the user should consider increasing these values to ensure finding the true global minimum for each site.
4. Although the algorithm first loops in time and then in space, this ordering can be reversed without any effect on the results obtained.
5. Since all sites are allowed to take a different number of neighbours, the heterogeneity of the solution depends exclusively on the measured data and thus no homogeneous or heterogeneous assumptions are taken by the user. This has the benefit of both limiting the user intervention and also allowing the data to identify the model best suited for the each application.
6. A significant computational advantage is obtained if the spatial dependence is limited to a number of sites smaller then the total number of measure-

Algorithm 1 Iterative Model Building

for $i = 1, 2, ...q$ **do**
  Order all sites in ascending order of distance relative to site s_i
  for $n = 1, 2, ...p_{max}$ **do**
    for $k = 1, 2, ...q_{max}$ **do**
      Estimate the first k elements of the i^{th} row of $A_1, A_2, ...A_n$, setting all other elements of the row to zero.
      Calculate the comparative predictive measure.
      Set the optimal temporal and spatial orders of s_i equal to the values of n and k giving the best comparative measure.
    end for
  end for
end for

ments being considered. Such an condition is common of many spatial studies as the pollution study being considered in this work.

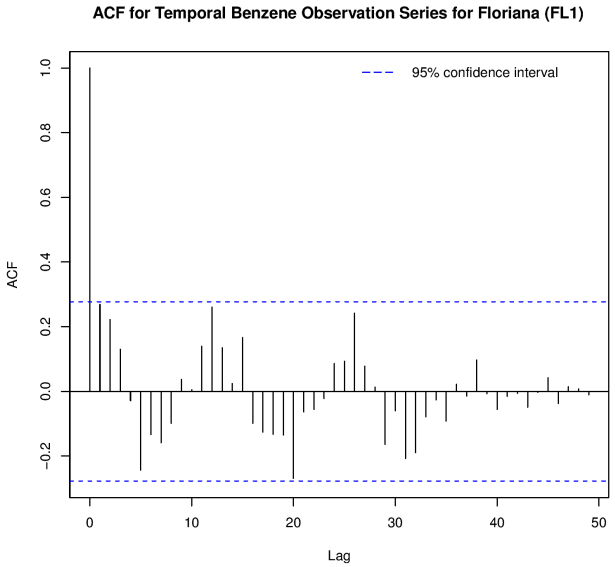

Figure 2: Autocorrelation Plot for the Benzene Observation Series at Floriana ('FL1').

4 An Example - Data-Driven Modelling of Air Pollution in Malta

To test the proposed methodology on a high-dimensional spatio-temporal application, a data-set of air pollution measurements gathered over the Maltese Islands for the period 2004 to 2010 will be used. This data is collected by the Malta Environment and Planning Authority (MEPA) using 123 passive diffusion tubes spread across the island, as shown in Figure 1. Three tubes are usually installed in each locality in sites categorized as near-road, intermediate or urban background. These tubes gather pollution levels for nitrogen dioxide, sulphur dioxide, ozone, benzene, toluene, xylene, ethyl benzene and o-xylene on a monthly basis. In this study, the pollutants usually associated with traffic, that is, nitrogen dioxide (NO_2) and benzene will be considered.

An initial analysis of the data indicated the presence of outliers and missing values. As typical of such studies (Barnett et al. 1994), outliers with measurements above $1.5 \times$ *Interquartile* Range+Upper Quartile or below 1.5×Interquartile Range-Lower Quartile were replaced by linear interpolations in time (Chatfield 2004). Missing values were also replaced by temporal linear interpolations. Note that such measures account only for 2% of the full dataset with respect to outliers and 4% with respect to missing values.

A typical temporal autocorrelation plot based on the pre-processed dataset is shown Figure 2. The low correlation values at each time delay indicate a short temporal dependence and thus point towards the choice of models with low temporal order. Similar plots for spatial correlations also reveal short distance spatial interactions and thus highlight the short range spatial dependency in the data. Partial correlation plots (rather then the full correlation plot of Figure 2), give similar indications. These chararteristics, common to various spatio-temporal applications, justify the need for methods that can efficiently deal with the spatial sparsity in the STAR model parameters.

The estimated models given by the methdology summarised in Algorithm 1 confirm these low-order dimensions for both the spatial and temporal domains. Specifically, for both benzene and nitrogen dioxide, a mean spatial order of 1.4 is obtained per site, for a total of approximately 172 parameters. Such results highlight a significant numerical advantage of approximately 99% in the number of estimated model parameters when comapared to the full model with (123 123) parameters for each temporal order. Figures 3 and 4 show AIC and BIC values for one of the sites situated in St Anne Street Floriana ('FL1') for benzene and nitrogen dioxide, re-

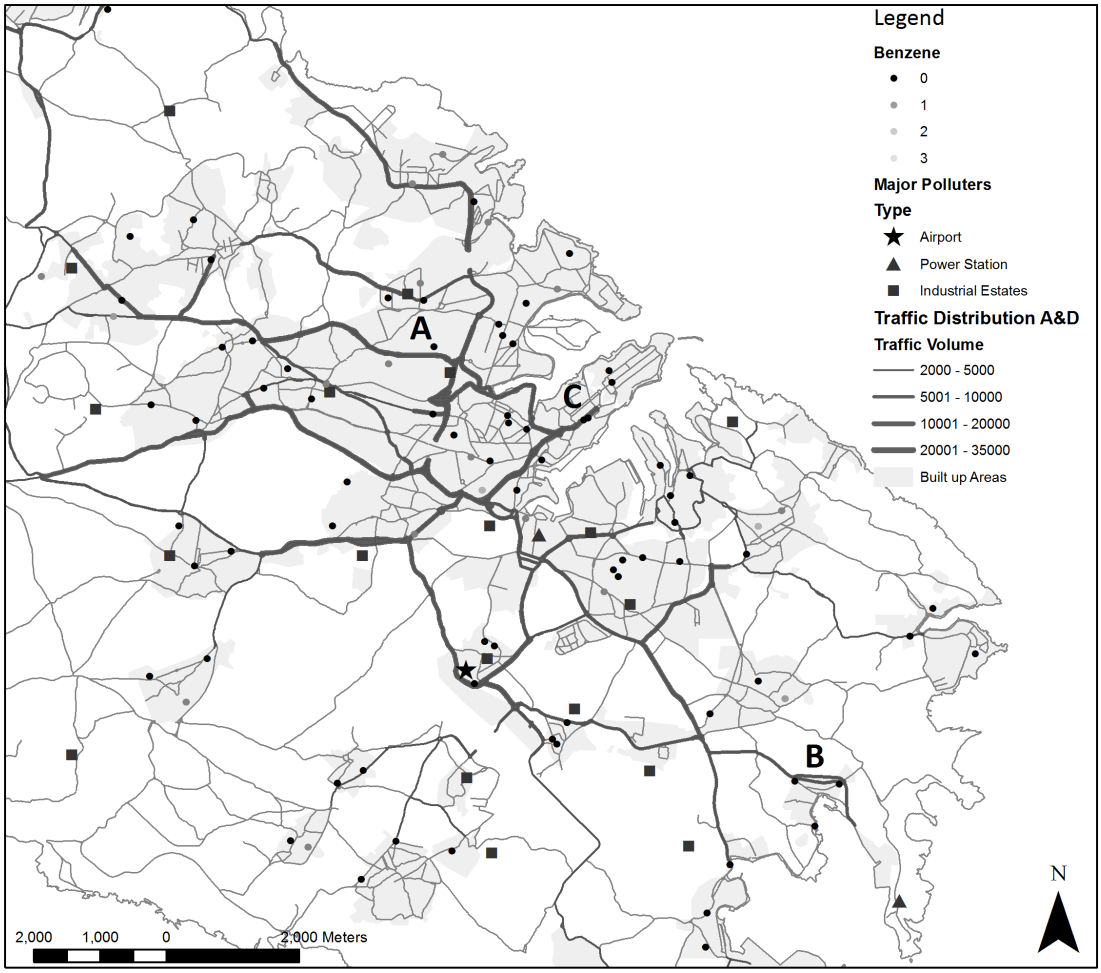

Figure 3: Number of Dependent Sites for Benzene.

spectively. Figure 3 indicates that benzene datapoints at Floriana can be described with a model of spatial order 1 and a temporal order 2, while Figure 4 indicates that nitrogen dioxide at the same site can be described using a model of a spatial and temporal order 1.

One step-ahead prediction estimates on a validation dataset (not used for estimation) of 12 months were used for model validation. For benzene datapoints the the Root Mean Squared Error (RMSE) for each monthly prediction has a mean of $1.661\,\mu\text{gm}^{-3}$ with a standard deviation of $0.664\,\mu\text{gm}^{-3}$, while for nitrogen dioxide datapoints the RMSE is $12.749\,\mu\text{gm}^{-3}$ with a standard deviation of $5.904\,\mu\text{gm}^{-3}$. These value represent 20% and 28% of the mean measurement respectively, and thus provide an acceptable predictor for the monthly pollutant concentrations. Moreover, the residues are spatio-temporally white up to a mean confidence interval of 98.7% for benzene and 95.4% for nitrogen dioxide, thus further confirming the validity of the predictions obtained.

5 Analysis of the Pollution Models

Air quality is of a major environmental concern in Malta as documented in several policy documents published over the past years (Government of Malta, 2002; Office of the Prime Minister, 2010). This concern, along with Malta's membership to the European Union in 2004, pose new obligation on the authorities to monitor the air quality. The main contributors to air pollution in Malta are the high demands for energy generation and the growth in private car use (NSO, 2010). The Maltese Islands were home to 229,016 private vehicles in 2009 (NSO, 2009), one of the highest car ownership rates in the world with approximately 669 cars per 1000 inhabitants. Such high vehicle ownership rates therefore highlight the need for continuous monitoring and analysis of the air pollutants mostly associated with traffic. The models obtained using the proposed methodology can thus be used to analyse local air pollution data and also

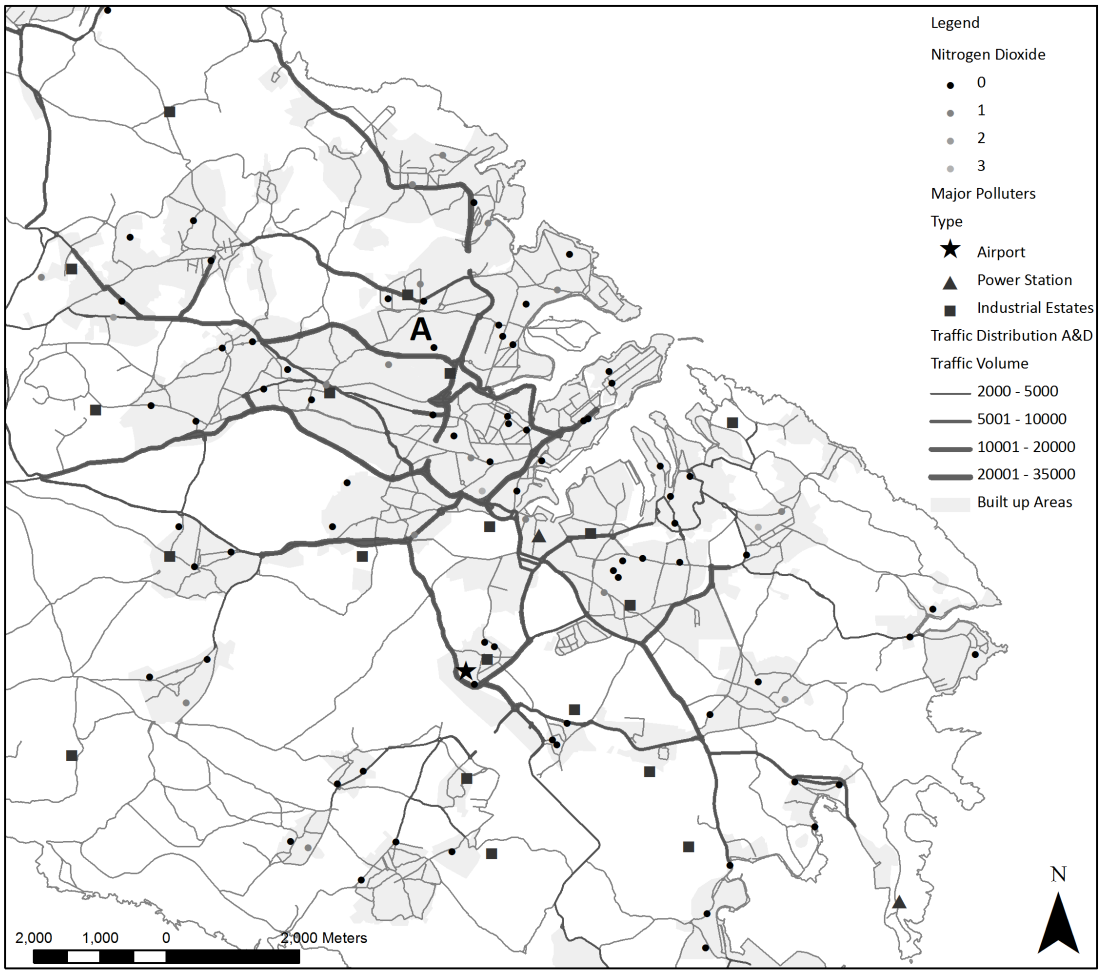

Figure 4: Number of Dependent Sites for $N0_2$.

evaluate the impact of future transport measures and possibly aid in the assesment of the health risks posed by air pollution.

Based on the models obtained, Figures 5 and 6 show the spatial model orders obtained for all observation sites for benzene and nitrogen dioxide, respectively. Note that, a value of 0 indicates that the reading at that particular site is only dependent on previous readings at the same site and a value of 1 indicates that the measurements are dependent on the site itself and its first closest neighbour, and so on for the other values. The low spatial orders observed in Figures 5 and 6 show that most sites are independent even though most of the modelled datapoints are located relatively close to each other and to pollution sources. Thus, the assumption that the dispersal of pollutants is equidistant and therefore one source of pollution in one area has an effect on the neighbouring areas is not supported by this data. This implies that the local pollution sources, rather than diffusion, have a predominant effect on a particular site. This is further highlighted by the inclusion in Figures 5 and 6 of potential sources of pollution in the main island such as traffic density, industrial estates, power stations and the airport.

The overall spatially independent behaviour of these pollutants would suggest that there are other, more local factors that are affecting air pollution. Some possible interpretations follow.

1. Since there is input from a stable source (such as traffic), similar temporal patterns can be observed. However, at different locations the source input levels may change (due to different traffic patterns) and therefore the behaviour of that point, even though it is relatively close, is independent. This is most evident in the area northwest of the Grand Harbour (marked A in Figures 5 and 6). This is reasonable since in the Maltese urban environment, the urban density, urban fabric and traffic, change considerably over a relatively short distance.
2. A few points experience higher spatial dependencies. These are marked with the letters B and C in Figure 5. In these cases we note that (i) the

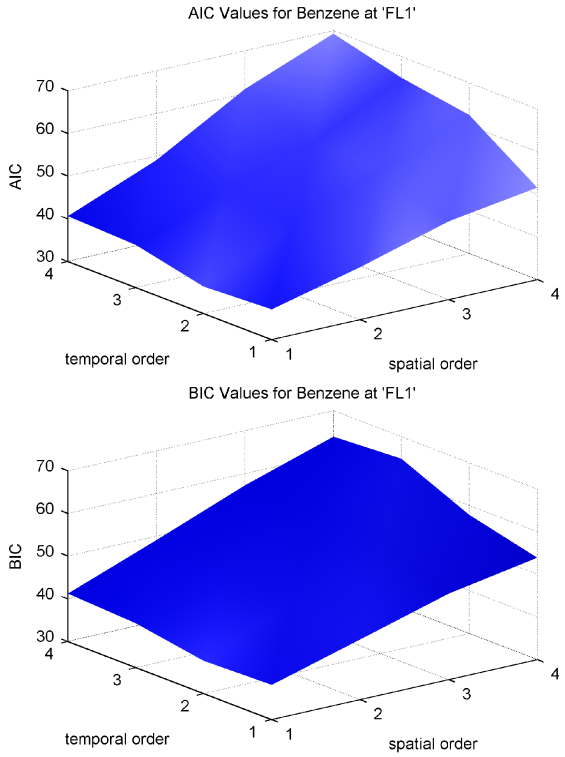

Figure 5: AIC and BIC Values for Benzene Datapoints at Floriana ('FL1').

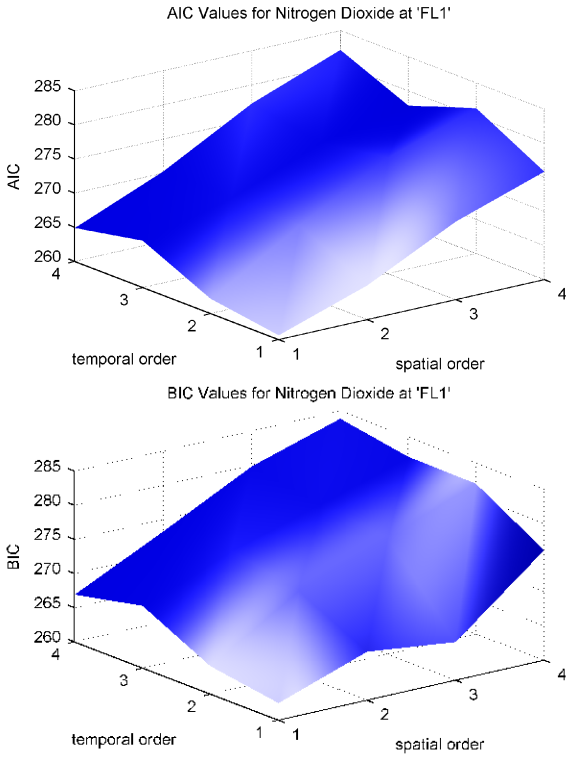

Figure 6: AIC and BIC Values for Nitrogen Dioxide Datapoints at Floriana ('FL1').

pollution values at some of these locations are relatively low, thus affecting the accuracy of the modelling procedure (area marked B) (ii) there are very similar environmental conditions (traffic and urban density) affecting these sites (area marked C).

3. These results are reasonable due to MEPA's approach in the identification of measurement sites. MEPA selects two to three sites per locality, one of which is a traffic site and the others are background sites without traffic. Thus, although geographically close, sites may exhibit significantly different traffic patterns and therefore different pollution measurements.

6 Conclusions

In this paper, a computationally efficient method for modelling heterogeneous spatio-temporal behaviour from large datasets was presented. This significant computational improvement was achieved through the use of the sparse spatial dependencies in the data. For the pollution measurement considered, a 99% reduction in the number of model parameters is obtained, resulting in a significant computational gain over classical estimation methods. In addition, one-step ahead predictions for air pollution concentrations performed on a validation dataset indicate estimation compatibilities comparable with classical methods.

Future work will focus on introducing measured pollution sources to the mathematical model to further examine the dependencies of the pollution readings on these sources. Alternative estimation techniques, such as the orthogonal least squares (Chen et al. 1989) and Expectation-Maximization (EM) algorithms (McLachlan et al. 2008), could also be used with the benefit of dealing with the estimation of the missing values in a more rigorous manner. In addition, one limitation to the available dataset is that only temporal dependencies over a monthly period can be captured due to the data's temporal resolution and thus shorter term dependencies cannot be ruled out. However the model and method presented can be readily applied to daily measurements when available, without any modifications. Such a finer temporal resolution has the added advantage of allowing for the introduction of exogenous inputs, such as wind and therefore avoiding the need to generalise such characteristics to monthly averages.

7 Acknowledgments

The research work disclosed in this paper is partly funded by the Malta Government Scholarship Scheme (MGSS).

References

Akinci A., D'Amico S., Malagnini L., Mercuri A., Akyol N. (2013) Scaling earthquake ground motion in the Western Anatolia, Turkey. *Physics and Chemistry of the Earth* (in press).

Azzaro, R. and Barbano, M.S. (2000) Analysis of the seismicity of Eastern Sicily: a proposed tectonic interpretation. *Ann. Geofis* 43(1), 171 - 188.

Beresnev, I. A., and Atkinson, G.M. (2002) Source parameters of earthquakes in eastern and western North America based on finite-fault modeling, Bull. *Seismol. Soc. Am.* 92, 695-710.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19, 716-723.

Amani A., Lebel T. (1997) Langrangian kriging for the estimation of Sahelian rainfall at small time steps. *Journal of Hydrology* 192, 125-157.

Barnett V., Lewis T. (1994) Outliers in Statistical Data. John Wiley and Sons, New York.

Berliner L.M., Wikle C.K., Cressie N. (2000) Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling. *Journal of climate. 13* 3953-3968.

Chatfield C. (2004) The Analysis of Time Series. Chapman and Hall / CRC, USA.

Chen S., Billings S.A., Luo W. (1989) Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control* 50(5) 1873 - 1896.

Chetcuti Zammit L., Scerri K., Attard M., Bajada T., Scerri M. (2011) Spatio-Temporal Analysis of Air Pollution Data in Malta. 11th International Conference of GeoComputation 2011, University College London, UK, 20th - 22nd July.

Cliff A.D., Ord J.K. (1974) Space-time modeling with applications to regional forecasting. *Transactions of the Institute of British Geographers.* 66, 119-128.

Cressie N.A.C. (1993) Statistics for Spatial Data. Wiley, NewYork.

Cressie N., Wikle C.K. (2011) Statistics for Spatio-Temporal Data. John Wiley and Sons, New Jersey.

De Luna X., Genton M.G. (2004) Spatio-Temporal Autoregressive Models for US Unemployment Rate. *Advances in Econometrics* 18, 279-294.

De Luna X., Genton M.G. (2005) Predictive Spatio-Temporal Models for spatially sparse environmental data. *Statistica Sinica* 15, 547-568.

Di Giacinto V. (2006) A generalized space-time ARMA model with an application to regional unemployment in Italy. *International Regional Science Review.* 29(2), 159-198. Government of Malta (2002). Johannesburg Summit 2002, Malta Country Profile.

Guo L.Z., Billings S. A. (2006) Identification of parial differential equations models for continuous spatio-temporal dynamical systems. *IEEE Transactions on Circuit and Systems - II.* 53(8), 657-661.

Ikegami T., Kaneko K. (1992) Evolution of host-parasitoid network through homeochaotic dynamics. *Chaos.* 2(3), 397-407.

Kessler D.A., Levine H., Reynolds W.N. (1990) Coupled-map lattice model of crystal growth. *Physical Review A.* 42(10), 6125-6128.

Leontaritis I.J., Billings S.A. (1985a) Input-output parametric models for non-linear systems. Part I: deterministic non-linear systems. *Int'l J of Control* 41, 303-328.

Leontaritis I.J., Billings S.A. (1985b). Input-output parametric models for non-linear systems. Part II: stochastic non-linear systems. *Int'l J of Control*41, 329-344.

Ljung L. (1999) System Identification - Theory For the User. Prentice Hall.

Lutkepohl H. (2005) New Introduction to Multiple Time Series Analysis. Springer-Verlag, Berlin.

Markos-Nikolus P., Martin-Gonzalez J. M. (2002) Spatial forecasting: Detecting determinism from single snapshots. *International Journal of Bifurcation and Chaos.* 12(2), 369-376.

Martin R. L. and Oeppen J. E. (1975). The identification of regional forecasting models using space-time correlation functions. *Transactions of the Institute of British Geographers.* 66, 95-118.

McCullagh P., Nelder J.A. (1989) Generalized Linear Models. Chapman and Hall, London.

MchLachlan G. J., Krishnan T. (2008) The EM Algorithm and Extensions. John Wiley and Sons, New York.

NSO (National Statistics Office) (2009). Motor Vehicles: Q4/2009 News Release. 20 January 2010. 008/2010. Available at http:// www.nso.gov.mt/statdoc/document_file.aspx?id=2669. [Accessed 19 January 2013].

NSO (National Statistics Office) (2010). Sustainable Development Indicators for Malta 2010. Available at http:// www.nso.gov.mt/statdoc/document_file.aspx?id=2913 [Accessed 19 January 2013].

Office of the Prime Minister (2010). Air Quality Plan for the Maltese Islands. Prepared by the Malta Environment and Planning Authority, Floriana, Malta. Available at http:// www.mepa.org.mt/airpublications. [Accessed 19 January 2013].

Peηa D., Tiao G. C., Tsay R. S. (2001) A course in Time Series Analysis. Wiley, New York.

Peng J., Wang P., Zhou N., Zhu J. (2009) Partial Corre-

lation Estimation by Joint Sparse Regression Models. *Journal of the American Statistical Association* 104:486, 735-746.

Pfeifer P. E., Deutsch S. J. (1980a) A three-stage iterative procedure for space-time modeling. *Technometrics.* 22(1), 35-47.

Pfeifer P. E., Deutsch S. J. (1980b) Identification and Interpretation of First-Order Space-Time ARMA Models. *Technometrics* 22(3), 397-403.

Pyle D. (1999) Data Preparation for Data Mining. Morgan Kaufmann Publishers Inc., San Francisco.Reiter C. A., (2005) A local cellular model for snow crystal growth. *Chaos, Solitons and Fractals.* 23, 1111-1119.

Schwarz, G. E. (1978) Estimating the dimension of a model. *Annals of Statistics* 6, 461-464.

Schwela D., Zali O. (1998) Urban Traffic Pollution. Spon Press, London.

Shen M., Chang G., Wang S., Beadle P.J. (2006) Nonlinear dynamics of EEG signals based on coupled network lattice model. *Advances in Neural Networks* 560-565.

Shen M., Lin L., Chang G. (2008) Novel coupled map lattice model for prediction of EEG signals. *Advances in Neural Networks* 347-356.

Shibata T., Kaneko K. (2002) Coupled map gas: structure formation and dynamics of interacting motile elements with internal dynamics. *Physica D.* 181, 197-214.