SPACE-TIME KERNELS

J.Q. Wang, T. Cheng*, J. Haworth

Department of Civil, Environmental and Geomatic Engineering, University College London,
Gower Street, WC1E 6BT London, United Kingdom {w.jiaqiu; tao.cheng; j.haworth}@ucl.ac.uk;

Commission II, WG II/3

KEY WORDS: Space-Time Kernels; Space-Time Analysis; Support Vector Regression;

ABSTRACT:

Kernel methods are a class of algorithms for pattern recognition. They play an important role in the current research area of spatial and temporal analysis since they are theoretically well-founded methods that show good performance in practice. Over the years, kernel methods have been applied to various fields including machine learning, statistical analysis, imaging processing, text categorization, handwriting recognition and many others. More recently, kernel-based methods have been introduced to spatial analysis and temporal analysis. However, how to define kernels for space-time analysis is still not clear. In the paper, we firstly review the relevant kernels for spatial and temporal analysis, then a space-time kernel function (STK) is presented based on the principle of convolution kernel for space-time analysis. Furthermore, the proposed space-time kernel function (STK) is applied to model space-time series using support vector regression algorithm. A case study is presented in which STK is used to predict China's annual average temperature. Experimental results reveal that the space-time kernel is an effective method for space-time analysis and modelling.

1. INTRODUCTION

Kernel methods are a class of algorithms for pattern recognition. The general task of pattern recognition is to find and study various patterns (such as clusters, correlations, classifications, regressions, etc) in different types of data (such as time series, spatial data, space-time series, vectors, images, etc) (Scholkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004). To date, kernel-based methods have been applied to a range of areas including machine learning and statistical analysis amongst others and have subsequently become a very active research area (Kanevski et al, 2009). Some of the best known algorithms capable of operating with kernels are support vector machines (Vapnik, 1995), general regression and probabilistic neural networks (Specht, 1991), canonical correlation analysis (Melzer et al, 2003), spectral clustering (Dhillon et al, 2004) and principal components analysis (Hoffmann, 2007).

Recently, kernel functions have been introduced to spatial analysis (Fotheringham et al, 2002; Hallin et al, 2004; Pozdnoukhov and Kanevski, 2008) and temporal analysis (Rüping, 2001; Ralaivola and d'Alché-Buc, 2004; Sivaramakrishnan et al, 2007). In the field of spatial analysis; Fotheringham et al (2002) developed a method using a Gaussian kernel function for the analysis of spatially varying relationships called Geographically Weighted Regression (GWR). GWR has been widely used for spatial analysis including house price prediction, ecological distribution, etc. Pozdnoukhov and Kanevski (2008) present a methodology for data modelling with semi-supervised kernel methods, which is applied to the domain of spatial environmental data modelling. They demonstrate how semi-supervised kernel methods can be applied in this domain, starting from feature selection; to model selection and up to visualization of the results. A case study of topo-climatic mapping reveals that the described methodology of data-driven modelling of complex environmental processes using machine learning methods improves the modelling considerably. In the field of temporal analysis, Ralaivola and d'Alché-Buc (2004) proposed a new kernel-based method as an extension to linear dynamical models. The kernel trick is used twice; first, to learn the parameters of the model, and second, to compute preimages of the time series predicted in the feature space by means of Support Vector Regression (SVR). Their model shows strong connection with the classic Kalman Filter model. Kernel-based dynamical modelling is tested against two benchmark time series and achieves high quality predictions. Sivaramakrishnan et al

(2007) propose a novel family of kernels for multivariate time-series classification problems. Each time-series is approximated by a linear combination of piecewise polynomial functions in a reproducing kernel Hilbert space by a novel kernel interpolation technique. Through the use of a kernel function, a large margin classification formulation is proposed, which can discriminate between two classes. The formulation leads to kernels, between two multivariate time-series, which can be efficiently computed. Furthermore, the proposed kernels have been successfully applied to writer independent handwritten character recognition.

The use of kernel methods in spatial and temporal analysis has been widely covered in the literature; however, how to accommodate kernels in spatio-temporal analysis is still unclear and hence forms the focus of the current study. The structure of the paper is as follows; in section two, a review of the relevant kernels that can be applied to spatial and temporal analysis is carried out; in section three; a space-time kernel (STK) function is proposed based on the principle of a convolution kernel that combines spatial and temporal kernels; in section four, a support vector regression machine is developed that makes use of STK (SVR-STK) to model space-time series. The final section summaries the major findings and proposes the direction of further research.

2. REVIEW OF KERNELS IN SPACE-TIME ANALYSIS

2.1 Kernels in spatial analysis

In spatial analysis, kernels are used as weighting functions to model and explain local spatial autocorrelation and heterogeneity features. For example, in Geographically Weighted Regression (GWR) (Fotheringham et al, 2002), a Gaussian kernel is used to model geographical data whose weights decrease continuously as the distance between the two points increases (note, Fotheringham et al (2002) also recommend the bi-square kernel function as an alternative). A Gaussian kernel, as seen in Figure 1, is defined as a symmetric monotonic function that decreases in value as the distance increases between the target spatial unit z_i and the neighbouring spatial unit z_j.

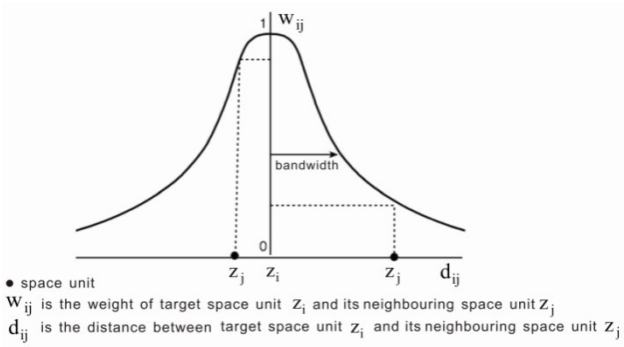


Figure 1. Sketch map of spatial kernel (Fotheringham et al, 2002)

The Gaussian kernel function takes the following form:

$$w_{ij} = \alpha \cdot e^{\frac{d_{ij}^2}{2\sigma^2}} \quad (1)$$

where d is the distance between target spatial unit z_i and its neighbouring spatial unit z_j and σ^2 is variance; also referred to as bandwidth (Fotheringham et al, 2002). The parameter σ^2 can change the smoothing degree of the Gaussian function curve; which alters the contribution of each neighbouring spatial unit z_j localized to a region nearby target spatial unit z_i. For a given regression point, the weight of a data point is at a maximum when it shares the same location as the regression point. This weight decreases continuously as the distance between the two points increases according to σ^2. In this way, a regression model is calibrated locally simply by moving the regression point across the region. For each location, the data will be weighted differently so that the results of any one calibration are unique to a particular location.

Kanevski et al (2009) apply a multi-scale kernel to deal with the problem of spatial interpolation of environmental data at different scales; the usual spatial interpolation methods are global and smoothing and can only deal with an average scale. This issue is addressed by considering a linear combination of Gaussian radial basis functions of different bandwidths. For a spatial modelling problem, multi-scale Radial Basis Functions (RBF) can be used:

$$f(x, a) = \sum_{i=1}^{N} \sum_{p=1}^{K} \left(a_i^{-} - a_i^{(p)}\right) e^{\frac{(x-x_i)^2}{2\sigma_p^2}} + b \quad (2)$$

where k is the number of kernels and $a_i^{(p)}$ is the weight corresponding to i -th training point and p -th kernel. A potential issue with this technique is that the choice of parameter k increases the dimension of the optimization problem, which is

$2N(k+1)$. Moreover, k and bandwidths σ_p have to be tuned, which can reflect the change of spatial process in scale.

2.2 Kernels in temporal analysis

Rüping (2001) provides an overview of some of the kernel functions that can be applied to time series analysis, and discusses their relative merits. Typically, time series analysis requires a higher level of reasoning than simple numerical analysis can provide and therefore model assumptions must be carefully considered. Experiments are carried out to discover if these different model assumptions have effects in practice and if kernel functions exist that allow time series data to be processed with support vector machines without intensive pre-processing. Rüping (2001) tests various kernel functions that are capable of being applied to time series analysis, including linear kernels, RBF kernels, Fourier kernels, Subsequence Kernels, PHMM Kernels, Polynomial kernels, etc. To give an example, a linear kernel $k(x,y)=x\cdot y$ is the most simple kernel function. The decision function takes the form $f(x)=w\cdot x+b$. When one uses the linear kernel to predict time series, i.e. $x_T=f(x_{T-1},\cdots,x_{T-k})=\sum_{i=1}^{k}w_i x_{T-i}+b$, the resulting model is a statistical autoregressive model of the order k ($AR[k]$). With the kernel, time series are taken to be similar if they are generated by the same AR-model.

Of most interest to this study is the Fourier kernel; since it can handle Fourier transformations. This representation is useful if the information of the time series does not lie in the individual values at each time point but in the frequency of some events. It was noted by Vapnik (1995) that the inner product of the Fourier expansion of two time series can be directly calculated by the regularized kernel function:

$$K_1(x,y)=\left\{\frac{1-q^2}{2(1-2q\cos(x-y)+q^2)}\middle|0<q<1\right\} \quad (3)$$

where q is regularization multiplier, which controls degree of attenuation of high frequency component in Fourier expanded equation. With the increase of q, SVR can express high frequency component more and enhance complexity of model. Conversely, with the reduction of q, high frequency component in data will attenuate quickly. Thus, the choice of q will influence the characterization ability of SVM for explaining the degree of data complexity. The schematic graph of Fourier kernel can be seen in Figure 2.

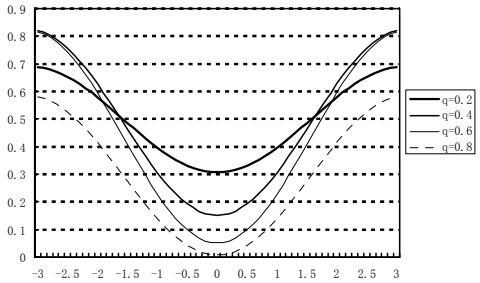


Figure 2. Schematic graph of Fourier kernel

3. SPACE-TIME KERNELS FUNCTION (STK)

The design of kernels for particular tasks is an open research problem. Kernel design methodology that incorporates prior knowledge into the kernel function is an important part of the successful application of the method (Kanevski et al, 2009). As discussed above, kernel functions can tackle spatial and temporal analysis using *kernel tricks* in machine learning and statistical models. The *kernel trick* is a method for using a linear classifier or regression algorithm to solve a nonlinear problem by mapping the original input space into a higher-dimensional feature space (Kanevski et al, 2009). According to kernel theory, a convolution kernel is a kind of construction kernel function, whose operation will be enclosed based on a standard kernel function (i.e. Polynomial kernel, Gaussian kernel, etc) (Haussler, 1999). A convolution kernel has following form:

$$K(x,y)=\sum_{x\in R^{-1},y\in R^{-1}}\prod_{i=1}^{J}K_i(x_i,y_i) \quad (4)$$

where R^{-1} is finite set and K is convolution of basic kernel functions $K_1,K_2,\cdots,K_D\ (K_1\times K_2\times\cdots\times K_j)$. We assume space-time kernel as $K_{ST}(x,y)$ and its form is:

$$K_{ST}(x,y)=\sum_{i=1}^{\lambda}\text{EMBED Equation.3}(K_S(x,y)\cdot K_T(x,y$$

(5)

where $K_{ST}(x,y)$ is space-time kernel, which processes space-time convolution; $K_S(x,y)$ is a spatial kernel, which processes spatial convolution; $K_T(x,y)$ is a temporal kernel, which processes temporal convolution; λ is the order of the kernel

function. Generally, a bigger λ can improve the learning ability of the kernel function. To avoid overfitting, λ should not be too large.

As discussed in Section 2.1, a Gaussian function is an important function that is able to tackle local spatial heterogeneous characteristics in geographical data. Additionally, Gaussian kernels have proven learning ability in machine learning regardless of the dimensionality of the sample data. Therefore, it can be used in the spatial kernel $K_S(x, y)$ discussed in Section 2.1 with following form:

$$K_S(x, y) = \left\{ \exp\left(-\frac{\|x - y\|^2}{\sigma^2} \right) \middle| \sigma > 0 \right\} \quad (6)$$

where $\|x - y\|$ is the distance between target spatial unit x and its neighbouring spatial unit; and σ^2 is the kernel bandwidth, which is a parameter for spatial kernel $K_S(x, y)$. σ^2 changes the smoothing degree of Gaussian curve, which varies the contribution of each neighbouring spatial unit x localized to a region nearby target spatial unit y .

Convolution theorem states that Fourier transformations can convert complex convolution operations to simple product operations (Nussbaumer, 1982). This indicates that Fourier kernels can be used to tackle convolution in time. The Fourier kernel has been discussed in Section 2.2. Additionally, it should be noted that the Fourier kernel is well suited to modelling periodic series (including *sine* and *cosine* frequency components). As for sequences there is no periodicity so a polynomial kernel is more appropriate due to its stronger generalization ability. A polynomial kernel takes the following form:

$$K_2(x, y) = \left\{ ((x \cdot y) + 1)^d \middle| d > 0 \right\} \quad (7)$$

where d is the order of the polynomial kernel. With reduction of d , generalization ability of the polynomial kernel will become stronger. Larger d will improve the complexity of the machine learning algorithm, resulting in the decline of generalization ability.

As discussed above, Fourier kernels and Polynomial kernels strongly complement each other. Therefore, we can combine them to approximate any series as long as kernel parameters are exact to the right degree. Thus, the temporal kernel $K_T(x, y)$ can be expressed mathematically as equation (8) where α is a coefficient to give more impact to the Fourier kernel K_1 and Polynomial kernel K_2 ; d and q are kernel parameters of the two kinds of basic kernel functions

According to Equation 5, 6 and 8, the expression of the space-time kernel can be derived as equation (9).

The function of Equation 9 is called the space-time kernel function (STK).

4. APPLICATION OF STK

To test the performance of STK, it is applied to the modelling of space-time series, which are sets of location-related time series (Bennett, 1975; Martin and Oeppen, 1975). The Support vector algorithm, one of the basic and most advanced algorithms, is a natural field of application for kernels. Hence, here an SVR model with STK is constructed and used to analyze and model nonlinear space-time series. Figure 3 describes the structure and target function of the SVR machine with STK. The output expression in Figure 3 is the objective function of SVR with STK (called SVR-STK) which is a regression function rather than a classification function.

.

$$\left(\alpha \cdot \frac{1 - q^2}{2(1 - 2q\cos(x - y + q^2)) + (1 - \alpha) \cdot ((x \cdot y) + 1)^d} \middle| 0 \le \alpha \le 1; d > 0; 0 < q < 1 \right\} \quad (8)$$

$$\left\{ \left(\exp\left(-\frac{\|x - y\|^2}{\sigma^2} \right) \right) \cdot \left(\alpha \cdot \frac{1 - q^2}{2(1 - 2q\cos(x - y + q^2)) + (1 - \alpha) \cdot ((x \cdot y) + 1)^d} \middle| \sigma > 0; 0 \le \alpha \le 1; d > 0; 0 < q < 1 \right) \right\} \quad (9)$$

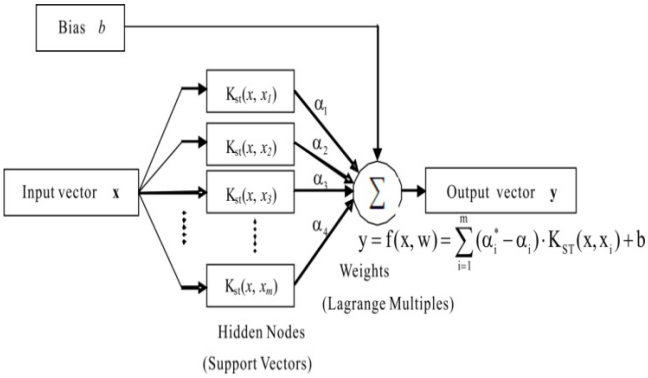
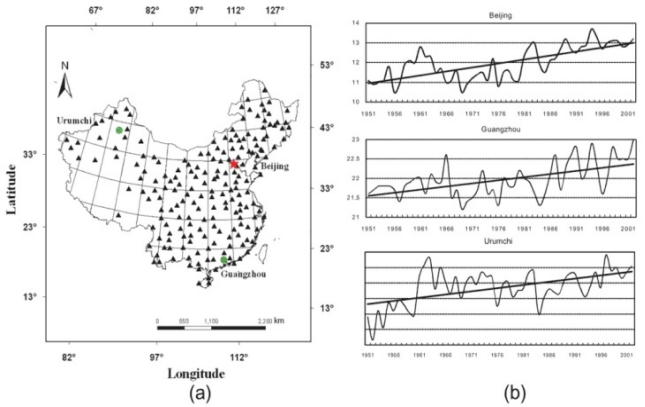


Figure 3. Architecture of support vector regression machine with space-time kernel (STK)

The model of Figure 3 is validated using data obtained from the national meteorological centre of P. R. China, including yearly temperature at 194 national meteorological stations (with geographical coordinates - longitude x and latitude y) from 1951-2002 as seen in Figure 4 (Cheng and Wang, 2009).



	Fitted (1951-1992)		
	RMSE		
	Plain SVR	Time series SVR	SVR-STK
Beijing	0.981	0.462	0.209
Guangzhou	0.910	0.314	0.084
Urumchi	1.173	0.853	0.306
	Forecasting (1993-2002)		
	RMSE		
	Plain SVR	Time series SVR	SVR-STK
Beijing	0.802	0.316	0.403
Guangzhou	0.813	0.418	0.387
Urumchi	0.837	0.551	0.541

Figure 4. Meteorological stations in study area: (a) spatial location distribution of the 194 stations; (b) graph of time series and trends of annual average temperature from 1951 to 2002 at the three stations of Beijing, Guangzhou, and Urumchi.

Of the 194 observation stations, there are huge data gaps in 57 stations. The data of these 57 stations are discarded, and data of 137 stations are used for the following test. To train and validate the models the data sets are split into two subsets: 80% as a sample set to train the model, and 20% as a validation set to test and validate the model. Thus, in this case, the meteorological data between 1951 and 1992 (42 years in total, nearly 80% of 52 years) is chosen as the training dataset for the forecasting between 1993 and 2002 (10 years in total, nearly 20% of 52 years).

Next, the SVR-STK model is constructed and trained after exploratory space-time analyses are undertaken. Each spatial unit is predicted in the experiment. Since the parameters of Equation 9 are numerous, selection of the arguments is tedious. The parameters of Equation 9 are adjusted and chosen according to the cross-validation method in order to obtain the best results. One-step-ahead forecasting, which is the most common testing standard, is considered in this case study. The SVR-STK results are compared firstly against a standard SVR model with inputs:

$$x_i, y_i, t_j \quad | \{i = 1, \ldots n, j = 1, \ldots m\} \tag{10}$$

Where x_i and y_i are the geographic coordinates of the i th station and t_j is the j th time period. Secondly, they are compared against pure time series SVR for the three individual test stations. The RBF kernel is used for both comparison tests; parameters were tuned separately for each station. Table 1 summarizes the accuracy measures using RMSE index for the fitted and forecasting results. SVR-STK significantly outperforms the plain SVR model for fitting and forecasting, achieving forecasting improvements of 49.75%, 52.4% and 35.36% for Beijing, Guangzhou and Urumchi respectively. SVR-STK also outperforms pure time series SVR for two of the three stations; Guangzhou and Urumchi, by 7.42% and 1.81% respectively. There is no improvement for Beijing, but given that SVR-STK requires only one set of parameters to be trained for all stations, the results are promising.

Table 1. Accuracy (RMSE) measures for three meteorological stations Beijing, Guangzhou and Urumchi in 52 years

5. CONCLUSIONS AND DISCUSSION

In the present paper, a space-time kernel function (STK) is presented, and the proposed STK is applied to the modelling of

space-time series by support vector regression algorithm. An illustrative case study is presented in which China's annual average temperature at 137 international meteorological stations from 1993-2002 is predicted using a support vector regression model with STK (SVR-STK). Although good results are achieved, further validation is still needed. Moreover, the following problems are identified; firstly, more research is needed into whether the proposed space-time kernel can be used to model and explain local space-time autocorrelation and heterogeneity, and secondly; whether the space-time kernel can be introduced to GWR modelling using some *kernel tricks*. The above two problems should be considered in further research.

Acknowledgements

This research is support by Chinese 973 Programme (2006CB701305) and 863 programme (2009AA12Z206), and UK EPSRC (EP/G023212/1).

References

Aizerman, M., Braverman, E., and Rozonoer, L., 1964. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25, pp. 821-837.

Bennett. R. J., 1975. The Representation and identification of spatio-temporal systems: an example of population diffusion in north-west England. *Transaction of the Institute of British Geographers*, 66, pp. 73-94.

Cheng, T., Wang, J.Q., 2009. Accommodating spatial associations in DRNN for space–time analysis. *Computers, Environment and Urban Systems*, 33(6), 409-418.

Dhillon, I.,Guan, Y., and Kulis, B. 2004. Kernel k-means, spectral clustering and normalized cuts. *In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, Seattle, WA, USA, pp, 551-556.

Fotheringham, S., Chris Brundson, A., and Charlton, M., 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley.

Hallin, M., Lu, Z., Tran, L.T., 2004. Kernel density estimation for spatial processes: the L1 theory. *Journal of Multivariate Analysis*, 88, pp. 61-75.

Haussler, D., 1999. *Convolution kernels on discrete structures*. Technical report, University of Santa Cruz.

Hoffmann, H., 2007. Kernel PCA for Novelty Detection. *Pattern Recognition*. 40. 863-874.

Kanevski, M., Pozdnoukhov, A., and Timonin, V., 2009. *Machine Learning for Spatial Environmental Data: Theory, Applications and Software*. EPFL Press.

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

Melzer, T., Reitera, M., and Bischof, H., 2003. Appearance models based on kernel canonical correlation analysis. *Pattern Recognition.* 36,pp, 1961-1971.

Nussbaumer, H. J., 1982. *Fast fourier transform and convolution algorithms*. Springer, Berlin.

Pozdnoukhov, A., and Kanevski, M., 2008. GeoKernels: modeling of spatial data on GeoManifolds. *In M. Verleysen, editor, ESANN 2008: European Symposium on Artificial Neural Networks – Advances in Computational Intelligence and Learning, Bruges*, Belgium, 23-25, April.

Ralaivola, L., and d'Alché-Buc F., 2004. Dynamical modeling with kernels for nonlinear time series prediction. *Advances in neural information processing systems*, 16, pp. 129 - 136.

Rüping, S., 2001. SVM kernels for time series analysis. In: R. Klinkenberg, S. Rüping, A. Fick, N. Henze, C. Herzog, R. Molitor, and O. Schröder (ed.), LLWA 01-Tagungsband der GI-Workshop-Woche Lernen-Lehren-Wissen-Adaptivitet, pp. 43-50.

Scholkopf, B., Smola, A., 2002. *Learning with kernel: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

Shawe-Taylor, J., Cristianini, N., 2004. *Kernel methods for Pattern Analysis*. Cambridge University Press.

Specht, D., 1991. A general regression neural network. *IEEE Transaction on Neural Network*. 2, pp. 568-576.

Sivaramakrishnan, K.R, Karthik, K., and Bhattacharyya, C., 2007. Kernels for large margin time-series classification. *IEEE Int Joint Conference on Neural Networks*. pp. 2746-2751.

Vapnik, V., 1995. *The nature of statistical learning theory*. New York, Springer-Verlag.