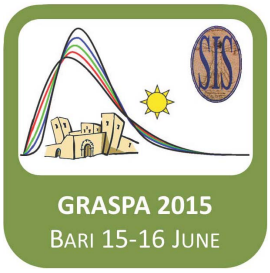

Patterns and Processes Revealed in High-Frequency Environmental Data

A. Elayouty[1,*], M. Scott[1], C. Miller[1] and S. Waldron[2]

[1] *School of Mathematics and Statistics, University of Glasgow, Glasgow, UK; a.el-ayouti.1@research.gla.ac.uk, Marian.Scott@glasgow.ac.uk, Claire.Miller@glasgow.ac.uk*
[2] *School of Geographical and Earth Sciences, University of Glasgow, Glasgow, UK; Susan.Waldron@glasgow.ac.uk*
[*] *Corresponding author*

Abstract. *High-frequency data are informative but also very challenging to analyze. Appropriate statistical tools are required to extract useful information from such data. A 15-minute resolution sensor-generated time series of the $EpCO_2$ from October 2003 to August 2007 in a small order river system in Scotland is used as an illustrative dataset. The aim of this paper is to study the daily patterns and dynamics of $EpCO_2$ using a Functional Data Analysis (FDA) approach. Using FDA, the discrete data within each day have been transformed to a smooth curve; then, a K-means clustering procedure has been applied to the spline coefficients defining the daily curves to identify the common daily patterns which can then be linked to underlying climatological and hydrological conditions.*

Keywords. *High-Frequency Data; Partial Pressure of Carbon Dioxide, Functional Data Analysis.*

1 Introduction

Advances in sensor technologies enable environmental monitoring programmes to record and store data at high temporal frequencies. These technical improvements in data acquisition present an opportunity to improve our understanding of environmental systems. However, to benefit from this wealth of data, appropriate statistical tools are required to manipulate and analyze large volumes of serially correlated data. In this paper, we consider a 15-minute resolution sensor-generated time series of the over-saturation of CO_2, $EpCO_2$, from October 2003 to August 2007 in a small order river system of the River Dee, Scotland. Surface waters are considered as key sources of atmospheric CO_2, therefore comprehensive understanding of the CO_2 dynamics in surface waters, quantified by the $EpCO_2$, is important. Due to the high-frequency nature of the data and the complex dynamics of $EpCO_2$ in relation to hydrodynamics, sophisticated exploratory tools and statistical models are needed to extract the main characteristics of the $EpCO_2$ series. One approach to analyze the high-resolution $EpCO_2$ time series is to investigate and model its variations and relationship with hydrology over time using wavelets and additive models (see [2] for details). Another strategy is to consider a functional data analysis approach, which is the main focus of this paper.

In Functional Data Analysis (FDA) [4], a time series can be treated as observations of a continuous function collected at finite series of time points, the observations of interest for data analysis are then curves over time. The paper describes the analysis of the daily dynamics of $EpCO_2$ using an FDA approach. In particular, the ultimate goal of the paper is to investigate the common daily patterns of $EpCO_2$ based on both mean level and shape, using functional clustering techniques. This, in turn, will help in determining the underlying climatological and hydrological conditions responsible for the different $EpCO_2$ daily patterns.

2 Methodology

In the context of FDA, the 96 (15-minute) observations within each day are considered as the discrete observations of a continuous smooth function. This view of the data allows the daily $EpCO_2$ patterns to be estimated using smooth curves removing issues of high correlations and variability between 15-minute observations. In this setting, the observations of interest are daily curves or functions, which are considered as realizations of a functional stochastic process $(X_i(t) : i \in \mathbb{Z})$, such that the time parameter i is discrete representing day of the year and the time parameter t is continuous representing time of the day. That is, $x_i(t)$ is regarded as the observation on day i, with intraday time parameter t.

Using the `fda` package in `R`, a smooth curve $x_i(t)$ is fitted for the observations within each day $(x_{i1},\ldots,x_{i96})$, $i = 1,\ldots,1095$ using cubic B-splines combined with a second-order roughness penalty, such that $x_i(t) = \sum_{k=1}^{K} a_{ik}\phi_k(t) = \mathbf{a}_i^T\Phi(t)$, where $\mathbf{a}_i$ is the vector of basis coefficients $(a_{i1},\ldots,a_{iK})^T$ to be estimated for the i^{th} sample path using penalized regression splines and $\Phi(t)$ is the vector of the basis functions $(\phi_1(t),\ldots,\phi_K(t))^T$.

With analogy to any classical statistical analysis, detecting outliers is crucial. Functional outliers can be identified using functional boxplots [5], developed based on the "band depth" measure which determines how deep or central a curve is. As with classical boxplots, functional boxplots are then constructed and functions are flagged as outliers if they fall outside the boxplot fences obtained by inflating the interquartile range (IQR) by $1.5 \times$ IQR. According to [5], functional boxplots are able to detect both shape and magnitude outliers (see [5] for more details).

After removing the detected outliers, a functional clustering procedure is performed to visualize the similarities and differences between the daily $EpCO_2$ curves and highlight the underlying climatological and hydrological conditions. One approach is to cluster the daily curves based on their spline coefficients using classical clustering techniques such as K-means [1]. The K-means procedure is iterative, in which the number of clusters is first specified, then each object is assigned to the cluster with the nearest mean such that the within-cluster sum of squares is minimized. The optimal number of clusters is initially selected using the gap statistic [6], which compares the change in the observed within-cluster dispersion with that expected under a null reference distribution of no clustering.

3 Results

Initially, a smooth curve is fitted for the observations within each day using saturated cubic B-splines combined with a roughness penalty. The smoothness of the curves is controlled by the smoothing parameter selected based on a sensitivity analysis. Next, functional boxplots were used to detect the out-

lying daily curves. The $EpCO_2$ curves of 29/8/2005, 23/9/2005 and 27/9/2005 are marked as shape and magnitude functional outliers and 26/9/2005 is flagged as a magnitude outlier. It is unsurprising to find outliers in adjacent days since the daily curves are time dependent. This number of potential outliers might decrease if the correlation between the curves is taken into account. However, 4 functional outliers represent only 0.4% of the total number of curves, and hence it was decided to delete these 4 curves before proceeding with any further analysis.

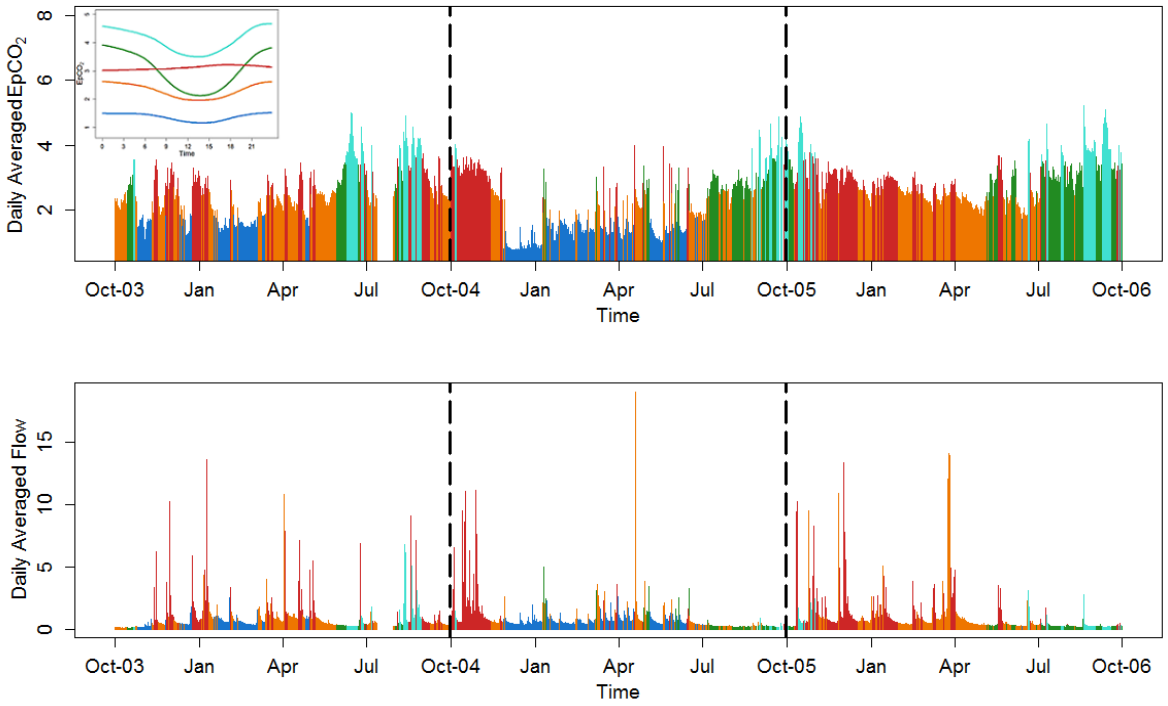

Figure 1: The average daily $EpCO_2$ (top) and flow (bottom) colored by their class membership obtained using K-means clustering of the $EpCO_2$ daily curves' splines coefficients. The top left legend shows the mean curve of each cluster.

After summarizing each daily curve with its estimated spline coefficients, K-means clustering was applied for a range of different number of clusters and the gap statistic was calculated to select the optimal number of clusters. The gap statistic identified consistently 5 optimal groups to represent the daily patterns of $EpCO_2$ for curves fitted using different smoothing parameters. This indicates that the optimal number of clusters is not sensitive to the chosen smoothing parameter. Next, a K-means procedure with 5 centers is applied to the spline coefficients. The top left legend in Figure 1 shows the grouping structure of the $EpCO_2$ daily curves which clearly depends on the estimated $EpCO_2$ mean levels. This is because the clusters are formed using the K-means procedure in which the classification is primarily based on the mean level. Another element of distinction between the 5 groups is the shape of the daily pattern of $EpCO_2$. Some of the groups have a clear daily cycle with a drop in the $EpCO_2$ during day time while others have a fairly constant $EpCO_2$ level over the day. The top and bottom panels of Figure 1 display the daily average $EpCO_2$ and flow (indicative for wet/dry days) respectively, and the class membership of each day according to the results of K-means clustering of the $EpCO_2$ daily curves. The figure shows that the turquoise and green curves representing a generally high $EpCO_2$ average with medium to severe drops in the $EpCO_2$ average level during the day light hours are more prominent in dry summer days, whilst the orange curves characterized by a lower $EpCO_2$ average and a medium trough during daytime underly the relatively wet spring and summer days. The blue curves with fairly stable and relatively

low levels of average $EpCO_2$ characterize the dry periods of winters and springs and the red group of curves consisting of a variable set of daily patterns often corresponds to the high flow events occurring in autumn and winter.

4 Discussion

In conclusion, FDA is shown to be a key tool in analyzing high-frequency environmental data. FDA has allowed the data observed every 15 minutes within each day to be expressed as continuous smooth functions without being concerned about the high-correlations between the 15-minute observations within the same day. After detecting the functional outliers, the primary results of functional clustering analysis indicated that the mean $EpCO_2$ level underlies the grouping structure. It is also evident that the $EpCO_2$ daily pattern is determined partly by the underlying hydrological (flow) and climatological (season) conditions.

Further work will investigate classifying the daily curves based on their Functional Principal Components scores (FPCs) using the classical clustering algorithms. Two key advantages for the use of FPCs are (i) identifying the primary sources of variations in the daily patterns of $EpCO_2$ and; (ii) the orthogonality and hence the independence between the FPCs of the same smooth curve. The shortcoming of either functional clustering approach described here is that the serial dependence between the daily curves has not been taken into account. Therefore, current work involves the extension of dynamic FPCs [3] which take advantage of the serial dependence between curves.

Acknowledgments. AE is grateful to the Glasgow University sensor studentship for funding.

References

[1] Abraham, C., Cornillon, P., Matzner-Lober, E. and Molinari, N. (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics* **50**, 581–595.

[2] Elayouty, A., Scott, M., Miller, C., Waldron, S. and Franco-villoria, M. (2015). Challenges in Modeling Detailed and Complex Environmental Data Sets: A Case Study Modeling the Excess Partial Pressure of Fluvial CO_2. *Journal of Environmental and Ecological Statistics* **Manuscript under revision**.

[3] Hormann, S. and Kidzinski, L. (2014) Dynamic Functional Principal Components. *Journal of Royal Statistical Society* **77**, 319–348.

[4] Ramsay, J.O. and Silverman, B.W. (1997). *Functional Data Analysis*. Springer.

[5] Sun, Y. and Genton, M. (2011). Functional Boxplots. *Journal of Computational and Graphical Statistics* **20(2)**, 316–334.

[6] Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of Royal Statistical Society* **63(2)**, 411–423.