Analysis and detection of outliers in water quality parameters from different automated monitoring stations in the Miño river basin (NW Spain)

J.I. Piñeiro Di Blasi a, J. Martínez Torres b, P.J. García Nieto c, J.R. Alonso Fernández d, C. Díaz Muñiz d, J. Taboada a,∗

a Departamento de Ingeniería de los Recursos Naturales y Medio Ambiente, Universidad de Vigo, 36310 Vigo, Spain
b Centro Universitario de la Defensa, Academia Militar, 50090 Zaragoza, Spain
c Department of Mathematics, Faculty of Sciences, University of Oviedo, 33007 Oviedo, Spain
d Cantabrian Basin Authority, Spanish Ministry of Agriculture, Food and Environment, 33071 Oviedo, Spain

A R T I C L E   I N F O
Article history:
Received 4 May 2013
Received in revised form 6 July 2013
Accepted 7 July 2013
Available online 15 August 2013

Keywords:
Functional data analysis
Outliers
Water quality monitoring
Monitoring samples
Functional depth

A B S T R A C T
Water quality controls help to prevent pollution and to protect public health as well as to maintain and improve the biological integrity of the water bodies, for which, authorities establish water quality standards. Water quality controls involve a large number of variables and observations, often subject to some outliers. An outlier is an observation that is numerically distant from the rest of the data or that appears strongly deviate from other members of the sample in which it occurs. Therefore, identification of atypical observations is an important concern in water quality monitoring and a difficult task because of the multivariate nature of water quality data. Our study provides a new method for detecting outliers in water quality monitoring parameters, using turbidity, conductivity and ammonium as indicator variables. Up to now, methods were based on considering the different parameters as a vector whose components were their concentration values. This innovative approach lies in considering water quality monitoring over time as continuous curves instead of as discrete points, i.e., the dataset studied is considered as a time-dependent function instead of as a set of discrete values in different time instants. This new methodology, which is based on the concept of functional depth, was applied to the detection of outliers in water quality monitoring samples in the Miño river basin with success. Results of this study are discussed here in terms of origin, causes, etc. Finally, the conclusions as well as the advantages of the functional method are exposed.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Human settlements are frequently located near the rivers, accentuating pollution problems with waste disposals from their urban and industrial centers. Water quality controls are necessary to avoid or minimize potential risks (García-Barcina et al., 2002; Rabalais et al., 2009). Monitoring is the key to understanding and managing complex ecosystems, and also for reporting both improvements and current status of water bodies to the Europe Union. Article 8 of the Directive 2000/60/EC, known as Water Framework Directive (WFD) (Directive 2000/60/EC, 2000), stresses the necessity to monitor surface water status. Its objective is to establish a coherent and comprehensive overview of water status within each river basin district. It must allow all surface water bodies to be classified into five classes: high, good, moderate, poor and bad.

The river Miño is the longest one in Galician region, and the largest once it receives the water from the river Sil, its main tributary. Specifically, 17,026 km² of Galician and Leonese regions are drained by these rivers. As a result, they come under strong pressures receiving wastewater from many cities and industries (i.e., Lugo, Ponferrada, Ourense, etc.). It is mandatory to intensify efforts in order to fulfill the requirements of the WFD (Directive 2000/60/EC, 2000). In this sense, the automated water quality monitoring is a very useful tool to check the status of water bodies (Andrews, 1984; Bartram and Rees, 2000; Aslan–Yilmaz et al., 2004) since it allows the real-time continuous monitoring of different parameters for the assessment of water quality.

Automated Water Quality Information System (AWQIS), implemented by Spanish Government between September 1993 and November 1995, can be considered a very useful tool for the
water status monitoring (Andrews, 1984; Bartram and Rees, 2000; Aslan-Yilmaz et al., 2004). AWQIS network currently consists of approximately 200 automatic alert stations spread all over the eight Spanish river basin districts. All of them are placed in areas with especially critical uses (water supply, protected areas, etc.) that need preventive actions or in points in which potential pollution episodes (large urban, industrial discharges, etc.) are expected. The main objective of this network is to record and transmit continuous information to the Ministry of Agriculture, Food and Environment and also to the data processing centers of the River Basin Authorities.

The pollutant controls are necessary to avoid or minimize potential risk (García-Barcina et al., 2002; Rabalais et al., 2009; García Nieto et al., 2013). It is reasonable to assume that some values of the potentially polluted water samples behave as outliers in a dataset (Hawkins et al., 2002; Dixit et al., 2005). Outliers are observations that substantially differ from the rest of the dataset. They can be classified as local extreme values (Saiz-Salinas and González-Oreja, 2000; Hastie et al., 2003) or global outliers. However, as the information is recorded every 5 min, the estimated data volume exceed 100,000 records a year for each AWQIS station. This means an additional difficulty to detect outliers.

Within the wide variety of data collected, this work is focused on turbidity, conductivity and ammonium values since these three variables are highly influential parameters in assessing water quality. Turbidity is a measurement of the suspended particulate matter and it is usually produced by anthropogenic sources as forest harvesting, road building, agriculture, urban developments, sewage treatment plant effluents, mining and industrial effluents (France and Peters, 1995; García Nieto et al., 2012; Alonso Fernández et al., 2013). Turbidity parameter is critical for water quality assessment since this decreases as its turbidity increases and vice versa. Furthermore, it is an integrative parameter because high turbidity values normally indicate high values of other parameters relative to water quality, such as chemical oxygen demand (COD), or ammonium, sulfate, or nitrate concentrations (Díaz Muñiz et al., 2012).

Besides the turbidity, the conductivity and ammonium concentration are important parameters to control and determine the status and the quality of water (Directive 2000/60/EC, 2000). On the one hand, it is well known that the ammonium concentration is strongly linked to the water quality because of its impact on the aquatic habitat. Examples of anthropogenic concentration sources are: sewage treatment plant effluent, urban developments and industrial effluents. Ammonium is one of the most important contributing factors to eutrophication in fresh water systems (Prepas et al., 2001). On the other hand, high conductivity values are linked to high salt concentrations (high salinity) (Akkyounlu and Akiner, 2012). Conductivity is the measure of water ability to conduct an electrical current. Nutrients, minerals, metals and any kind of pollution can affect this ability (Doering, 1996; Clark, 2001; Boesch, 2002). Consequently, conductivity is a parameter strongly linked to human activities.

The aim of this research work is to build a model to identify temporal outliers in samples from the water quality monitoring in the Miño river basin (17,026 km²) where there are 5 AWQIS stations to monitor the water quality of this river (see Fig. 1).

In this study, the functional method of data analysis was applied successfully. This new functional approach allows treating the dataset as continuous measurements, that is to say, measurements over time without discrete values. Therefore, this method provides functional outliers, allowing to analyze the trend and periodicity of the measurements, without regard as outliers those discrete measurement errors. Furthermore, functional approach has the advantage with respect to most of the criteria for outliers detection based on a discrete approach that it does not require a data set normally distributed. Therefore, an additional processing is not necessary.

This innovative research work is structured as follows. In the first place, the necessary materials and methods are described to carry out this study. Next the obtained results are shown and discussed. Finally, the main conclusions drawn from the results are exposed.

2. Materials and methods

2.1. Study area and dataset

The river Miño is the most important one in the Galician region (Northwestern Spain). The source of the Miño lies about 50 km north of Lugo, in Galician region. Twenty kilometers north of Orense at Os Pareas, the river Miño receives the waters of its main tributary, the river Sil. Passing Orense, the river Miño flows in a southwest direction until reaching Portugal. It borders between Spain and Portugal, along about 80 km before reaching the Atlantic Ocean. Its length is about 310 km and its annual mean flow is 340 m³/s. The river Sil, its main tributary, is a little bit shorter (228 km long) and its annual mean flow is 100 m³/s.

The Miño-Sil Basin Authority (from the Spanish Ministry of Agriculture, Food and Environment) has 5 automated monitoring stations (see Fig. 1): three stations are located on the river Miño (Stations 107, 108 and 109) and the other two stations on the river Sil (Stations 110 and 123).

Regarding the climate, two main climatic zones can be distinguished in Spain: the Mediterranean and the oceanic climate (Pausas, 2004). Overall, the Galician region has a mild climate influenced by the ocean so it is very rainy. Nevertheless, its irregular topography results in multiple microclimates with strong variations in areas of about 200 km²: corresponding to the part of the river Miño which borders between Spain and Portugal, with a climate very close to the mild Mediterranean (Johnson et al., 2000). The Galician region has an average annual temperature of 13.3 °C (8.5 °C in winter time and 19 °C in summer). The land corresponding to the river Sil and its tributaries has an oceanic climate with abundant rainfalls (between 1200 and 2000 mm), in combination with a strong Mediterranean influence giving place to the decrease these rainfalls during summer time (Pausas, 2004).

The data used for the analysis of functional outliers are the values continuously collected from 2007 to 2010, of the following physical–chemical parameters:

- **Turbidity**: It is the cloudiness or haziness of the water caused by individual particles (suspended solids) that are generally invisible to the naked eye, similar to smoke in air. The measurement of turbidity is a key test of water quality. Turbidity in open water may be caused by growth of phytoplankton. Human activities that disturb land, such as construction, mining and agriculture, can lead to high sediment levels entering water bodies during rain storms due to storm water runoff. Areas prone to high bank erosion rates as well as urbanized areas also contribute large amounts of turbidity to nearby waters, through stormwater pollution from paved surfaces such as roads, bridges and parking lots. Certain industries such as quarrying, mining and coal recovery can generate very high levels of turbidity from colloidal rock particles. Therefore, turbidity is in relation to water transparency.

- **Temperature, conductivity and pH**: Relative to the water body stratification.

- **Dissolved oxygen**: It is a relative measure of the amount of oxygen that is dissolved or carried in a given medium. It can be measured with a dissolved oxygen probe such as an oxygen sensor in the
water. Oxygen gets into water by diffusion from the surrounding air, by aeration (rapid movement), and as a waste product of photosynthesis. Total dissolved gas concentrations in water should not exceed 110 percent. Concentrations above this level can be harmful to aquatic life. Furthermore, adequate dissolved oxygen is necessary for good water quality. Oxygen levels that remain below 1–2 mg/l for a few hours can result in large fish kills.

- **Ammonium**: This cation is a positively charged polyatomic ion with the chemical formula \( \text{NH}_4^+ \). It is formed by the protonation of ammonia (\( \text{NH}_3 \)). Ammonium ions are a waste product of the metabolism of animals. In fish and aquatic invertebrates, it is excreted directly into the water. Ammonia–nitrogen is an inorganic, dissolved form of nitrogen that can be found in water and is the preferred form for algae and plant growth. Ammonia is the most reduced form of nitrogen and is found in water where dissolved oxygen is lacking. It is in relation to nutrients content.

The data set was continuously collected at the above-mentioned five stations from 2007 to 2010. The automated water quality monitoring stations collect data 24 h a day, 7 days a week. Measurements of water quality parameters are taken every 15 min. Next they are combined and transformed into hourly averages using data loggers in situ.

In summary, this study is focused solely on the analysis of the turbidity, conductivity and ammonium among all the parameters collected in these automated monitoring stations. The sampling was carried out at 75 cm below the water surface. The sensors used for measuring the studied parameters at the automated monitoring stations were:

- **Turbidity**: This parameter was measured using the immersion sensor CUS31 Turbimax W, both for drinking water and industrial water. Its measurement principle is an optical sensor 90° from scattered light to near infrared light (880 nm).
- **Conductivity**: This parameter was measured by the Condumax W CLS21 sensor made up of two electrodes with a fixed cable. Its measurement principle is a conductive cell with graphite electrodes for midrange applications.
- **Ammonium**: This variable was measured through a CAS40D ISE-max device, made up of a sensor of electrodes and a transmitter (CM442). It performs the measurement using an ion selective electrode for the continuous measurement of the ammonium.

### 2.2. Constructing curves from points: smoothing

The first step to solve the problem with the proposed method consists on generating the functional sample from the vector sample, i.e. to build best-fitting curves with the points corresponding to the discrete values from the experimental measurements. Thus, we do not work with the set of observations as multivalued vectors, but with a set of observations considered as continuous functions over time (Ramsay et al., 2009).

In this way, functional data are observations of a continuous random process observed at discrete points (Ramsay et al., 2009; Ramsay and Silverman, 2010; Martinez Torres et al., 2011; Díaz Muñiz et al., 2012). In consequence, given a set of observations \( x(t_j) \) in a set of \( n_p \) points \( t_j \in \mathbb{R} \), where \( t_j \) represents each time step, and \( n_p \) represents the number of observations. All the observations can be considered as discrete observations of the function \( x(t) \in \mathcal{X} \subset F \), where \( F \) is a functional space. In this regard, to estimate the function \( x(t) \), it is considered that \( F = \text{span} \{ \phi_1, \ldots, \phi_k \} \), where \( \{ \phi_k \} \), with \( k = 1, 2, \ldots, n_b \), is a set of basis functions, \( n_b \) the number of basis functions necessary to generate the basis of the functional space. This set of basis functions can be of different types, but it is common to use spline or Fourier functions. For convenience, we consider this
expansion (Alameddine et al., 2010; Ramsay et al., 2009; Martínez Torres et al., 2011; Díaz Muñiz et al., 2012):

\[ x(t) = \sum_{k=1}^{n_b} c_k \phi_k(t) \]  

(1)

where \( \{c_k\}_{k=1}^{n_b} \) represents the coefficients of the function \( x(t) \) with respect to the chosen set of basis functions. At this point, it is possible to conclude that the smoothing problem is to solve the following regularization problem (Martínez Torres et al., 2011; Díaz Muñiz et al., 2012):

\[ \min_{x \in F} \sum_{j=1}^{n_b} (x_j - x(t_j))^2 + \lambda \Gamma(x) \]  

(2)

where \( x_j = x(t_j) + \epsilon_j \) (with \( \epsilon_j \) as the value of the zero-mean random noise) is the result of the observation x point \( t_j \). \( \lambda \) is a regularization parameter that regulates the intensity of the regularization and \( \Gamma \) is an operator that penalizes the complexity of the solution. Taking into account the expansion in Eq. (1), the previous problem can be written as (Ramsay et al., 2009; Díaz Muñiz et al., 2012):

\[ \min_{c} \{ (z - \Phi c)^T (z - \Phi c) + \lambda \Phi c^T R \Phi c \} \]  

(3)

where \( z = (z_1, \ldots, z_{n_b})^T \) is the observation vector, \( c = (c_1, \ldots, c_{n_b})^T \) is the vector of coefficients of the functional expansion, \( \Phi \) is the \( n_p \times n_b \) matrix of elements \( \Phi_{k} = \phi_k(t_j) \), and \( R \) is the \( n_b \times n_b \) matrix of elements (Ramsay et al., 2009):

\[ R = (D^2 \phi_k, D^2 \phi_l)_{L^2(T)} = \int_T D^2 \phi_k(t) D^2 \phi_l(t) \, dt \]  

(4)

where \( D^2 \phi(t) \) represents the nth-order differential operator of the function \( \phi_k \). Clearly, the solution of this problem can be calculated as follows:

\[ c = (\Phi^T \Phi + \lambda R)^{-1} \Phi^T z \]  

(5)

2.3. The functional depth concept

It is well-known that the depth measurement was originally introduced in the multivariate analysis to measure the centrality of a point in relation to a cloud of points. In other words, the depth provides a way to arrange the points in a Euclidean space from the center to the periphery, so that the points closer to the center will have a greater depth. In this sense, the concept of depth, has recently been extended to functional data (Cuevas et al., 2006; Ramsay et al., 2009; Martínez Torres et al., 2011; Díaz Muñiz et al., 2012). Therefore, the functional depth measures the centrality of a curve \( x_i \) within a set of curves \( x_1, \ldots, x_n \).

Summarizing, the most important measurements of functional depth are as follows:

- **Fraiman–Muniz depth (FMD):** Let \( F_{n,t}(x_i(t)) \) be the cumulative empirical distribution function (Fraiman and Muñiz, 2001) for the values of the curves \( |x_i(t)|_{e=t}^{n} \) in a time \( t \in [a, b] \) governed by (Díaz Muñiz et al., 2012):

\[ F_{n,t}(x_i(t)) = \frac{1}{n} \sum_{k=1}^{n} I(x_k(t) \leq x_i(t)) \]  

(6)

where \( I(\cdot) \) is the indicator function. The FMD for a curve \( x_i \) with respect to set \( x_1, \ldots, x_n \) is given by:

\[ \text{FMD}_n(x_i(t)) = \int_a^b D_n(x_i(t)) \, dt \]  

(7)

being \( D_n(x_i(t)) \) is the depth of the point \( x_i(t), \forall t \in [a, b] \) given simply as:

\[ D_n(x_i(t)) = 1 - \frac{1}{2} - F_{n,t}(x_i(t)) \]  

(8)

- **H-modal depth (HMD):** It is important to note that the functional mode (based on the mode concept) is defined as the curve more densely surrounded by the other curves in the sample. H-modal depth (HMD) is written as (Cuevas et al., 2006; Ramsay et al., 2009; Martínez Torres et al., 2011; Díaz Muñiz et al., 2012):

\[ \text{HMD}_n(x_i, h) = \sum_{k=1}^{n} K \left( \frac{|x_i - x_k|}{h} \right) \]  

(9)

so that \( K : \mathbb{R}^+ \rightarrow \mathbb{R}^+ \) is a kernel function, \( ||\cdot|| \) is a norm in a functional space and \( h \) is the bandwidth parameter. One of the most widely used norms for a functional space is \( L^2 \) is given as follows (Ramsay et al., 2009):

\[ ||x_i(t) - x_j(t)||_{L^2} = \left( \int_a^b (x_i(t) - x_j(t))^2 \, dt \right)^{1/2} \]  

(10)

Furthermore, the infinite norm \( L^\infty \) is used sometimes, leading to:

\[ ||x_i(t) - x_j(t)||_{L^\infty} = \sup_{t \in [a,b]} |x_i(t) - x_j(t)| \]  

(11)

Conceptually, different kernel functions \( K(\cdot) \) can also be defined, among them the truncated Gaussian kernel, which is used in this research work due to its high efficiency in detecting functional outliers (Cuevas et al., 2006; Martínez Torres et al., 2011; Díaz Muñiz et al., 2012):

\[ K(t) = \frac{2}{\sqrt{2\pi}} \exp \left( -\frac{t^2}{2} \right), \quad t > 0 \]  

(12)

2.4. Functional outliers

A set of functional samples may contain elements that, despite not incorporating errors in themselves, may have different patterns. Depth measurements described above, used to identify outlier functional samples in functional samples, allow sets of observations over time fitted to curves to be compared, rather than only the mean values in the measurement time interval.

Depth and outlier are two inverse concepts. Indeed, an outlier for a functional sample will have considerably less depth. Similarly, the curves with the greatest depths are sought in order to identify functional outliers. In this sense, H-modal depth (HMD) was used to generate the selection criterion of outliers, selecting the value of bandwidth \( h \) as the 15th percentile of the empirical distribution \( \{||x_i - x_j|||, i, j = 1, \ldots, n\} \), because the model is not substantially sensitive to the choice of the parameter \( h \) (we have checked various values such as 10 and 20, and the results were equal). Furthermore, we have selected the standard bandwidth value according to the parents of the functional outliers methodology (Cuevas and Fraiman, 1997). Additionally, the cut-off value \( C \) was selected in such a way that the percentage of correct observations misidentified as outliers (type I error) was approximately 1%, according to (Febrero-Bande et al., 2008). Therefore, we have that:

\[ \text{Pr}(|\text{MD}_n(x_i(t)| \leq C)) = 0.01, \quad i = 1, \ldots, n \]  

(13)

It is important to point out that, unfortunately, the distribution of the chosen functional depth is not known a priori, requiring that the value for \( C \) had to be estimated. Note that of the different methods to estimating this value (Cuevas et al., 2006; Martínez Torres et al., 2011).
et al., 2011; Díaz Muñiz et al., 2012), we have chosen, for the purpose of this research, a technique based on bootstrapping (Cuevas et al., 2006; Febrero-Bande et al., 2007, 2008; Peng and Qi, 2008; Ramsay et al., 2009; Martínez Torres et al., 2011) of the curves of the original set with a probability proportional to depth. In this way, the bootstrapping approach can be summarized as follows (Díaz Muñiz et al., 2012):

- Firstly, a new sample is extracted from the original sample by means of sampling with replacement (in other words, each extracted element is replaced after extraction and so it may be selected again). Furthermore, a resampling of order 10 has been selected.
- Secondly, based on this new sample, the populational parameter of interest is derived on the basis of the construction of a statistic.
- Thirdly, using this result, the two steps above are repeated until a large number of estimates are obtained.
- Finally, the empirical distribution of the statistic, which estimates the cut-off value $C$, is determined.

3. Analysis of results and discussion

For the three studied variables in this research work, the sample depended on each of the components analyzed. Our sample, $(x_{i})_{j=1}^{48}$, corresponds to 47 months from January 2007 to November 2010, where $x_{i}$ was the concentration of each variable measured on the day $i$ of the month $j$, with $i = 1, 2, ..., 30$ and $j = 1, 2, ..., 48$. After applying the smoothing exactly as it was described above, a new sample for each of the three analyzed variables was obtained, where each $x_{i}$ is now a function of basis functions with 1000 elements. This calculation gives place to a correlation coefficient of 0.95 between the discrete values and the observations of each function in the corresponding points. In other words, the generated functional sample is correlated with the discrete sample by 95%. This correlation can be observed in Fig. 2 for the turbidity with an example (May 2010) of the function registered for the turbidity and the initial data (note that there is 2880 points for each month corresponding to a 15 min observation interval). Once we have constructed the sample by the functional smoothing process, an analysis of the functional depths was carried out.

![Turbidity function smoothed](image)

**Fig. 2.** Turbidity function registered (solid line) after applying the smoothing process along with initial data (crosses).

### Table 1

<table>
<thead>
<tr>
<th>Station</th>
<th>Year</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>107</td>
<td>$6, 11$</td>
<td>–</td>
<td>2, 3, 11</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>108</td>
<td>1, 6, 7</td>
<td>9, 10</td>
<td>2, 3, 4</td>
<td>1, 10, 11</td>
<td></td>
</tr>
<tr>
<td>109</td>
<td>1, 3, 6, 7, $11$</td>
<td>1, 2, 4</td>
<td>2, 3</td>
<td>1, 11</td>
<td></td>
</tr>
<tr>
<td>110</td>
<td>3, 6, $11$</td>
<td>1, 9, 10</td>
<td>2</td>
<td>3, 11</td>
<td></td>
</tr>
<tr>
<td>123</td>
<td>$6, 11$</td>
<td>5, 10</td>
<td>2</td>
<td>11</td>
<td></td>
</tr>
</tbody>
</table>

Note: The numbers indicate the month of year (i.e., 1 is January, 2 is February and so on) as well as the kind of outlier: positive outliers in normal font and negative outliers in bold italic font.

After creating a functional sample, we have proceeded to the analysis of functional depths. After its application and analysis, a complete match among the outliers detected for the three parameters (turbidity, ammonium and conductivity) are observed in several months of 2007, 2008, 2009 and 2010.

Functional analysis used in this study has several advantages over classical vector analyses (Díaz Muñiz et al., 2012). In practice, the functional analysis does not require a Gaussian initial dataset or to subject the non-Gaussian initial dataset to statistical Box–Cox power transformations. In this sense, only the functional method has been performed here.

Weather conditions can contribute to outliers presence since they influence both flow and composition of the river. As it can be observed in Fig. 3(a) and (b), and according to the expectations, there is a strong correlation between rainfall in a specific area and the river flow in this same area.

Rainfalls give place to two opposite effects in the composition of river water. On the one hand, rainfalls cause the dilution of substances into the river. On the other hand, they also drag substances from the adjacent lands so that:

- If there is ammonium in this dragging (i.e., from the fertilizers of meadows), the ammonium concentration is increased into the river.
- If the dragging contains salts, the conductivity is also increased.
- The turbidity is always increased taking into account that a dragging contains abundant insoluble substances.

Indeed, most of the positive turbidity outliers (see Table 1) coincide with time periods where the river flow reaches a local maximum abruptly suggesting that they are due to a significant dragging of materials along the river (washing of substances into the river) caused by heavy rain (see Fig. 3(b) and (b)). On the contrary, negative turbidity outliers (see Table 1) coincide with time periods where the river flow is low. Exceptions can be observed for some outliers detected in 2008 (Stations 110 and 123). Perhaps, there is no a significant washing of materials into the river or other

### Table 2

<table>
<thead>
<tr>
<th>Station</th>
<th>Year</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>107</td>
<td>2, 3, 4, 5, 6, 7</td>
<td>2, 4</td>
<td>10, 11, 12</td>
<td>5, 6</td>
<td></td>
</tr>
<tr>
<td>108</td>
<td>3, 5, 6, 9, $11$</td>
<td>1, 2, 4, 8, 9</td>
<td>5</td>
<td>6, 9</td>
<td></td>
</tr>
<tr>
<td>109</td>
<td>4, 5, 6, 8, 9, 10</td>
<td>–</td>
<td>7, 8</td>
<td>8, 9</td>
<td></td>
</tr>
<tr>
<td>110</td>
<td>1</td>
<td>–</td>
<td>9, 10, 11</td>
<td>13, 4</td>
<td></td>
</tr>
<tr>
<td>123</td>
<td>2, 5, 6, 9, $11$</td>
<td>1, 2, 5, 10</td>
<td>4, 5</td>
<td>9, 11</td>
<td></td>
</tr>
</tbody>
</table>

Note: The numbers indicate the month of year (i.e., 1 is January, 2 is February and so on) as well as the kind of outlier: positive outliers in normal font and negative outliers in bold italic font.
Fig. 3. (a) Meteorological data: average monthly precipitation in Lugo, Ponferrada, Orense and Vigo; and (b) river flow in cubic meters per second as a function of time from January 2007 to November 2010.

Table 3

<table>
<thead>
<tr>
<th>Station</th>
<th>Year</th>
<th>Year</th>
<th>Year</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>107</td>
<td>6, 7, 11</td>
<td>–</td>
<td>2, 3</td>
<td>7, 11</td>
</tr>
<tr>
<td>108</td>
<td>1, 7</td>
<td>–</td>
<td>1, 3</td>
<td>7, 8, 11</td>
</tr>
<tr>
<td>109</td>
<td>1, 6, 9, 11</td>
<td>1, 2, 9, 10, 11</td>
<td>2, 3</td>
<td>8, 9, 11</td>
</tr>
<tr>
<td>110</td>
<td>10, 11</td>
<td>–</td>
<td>–</td>
<td>10, 11</td>
</tr>
<tr>
<td>123</td>
<td>9, 10, 11</td>
<td>–</td>
<td>3, 12</td>
<td>9, 10, 11</td>
</tr>
</tbody>
</table>

Note: The numbers indicate the month of year (i.e., 1 is January, 2 is February and so on) as well as the kind of outlier: positive outliers in normal font and negative outliers in bold italic font.

4. Conclusions

Several quality water data for the Miño river basin (Northwestern Spain) have been analyzed to detect outliers with success. The factors (i.e., some clean discharges without significant increases of turbidity) have stronger influence in these sites at that time.

Tables 2 and 3 show the outliers corresponding to ammonium and conductivity variables. It is possible to observe that there is no correlation with rainfalls and the river flow because of the two above-mentioned opposite effects of the rainfalls on soluble substances such as ammonium or salts in general.
method is based on treating the point data as functions. This technique leads to a better information recovery from the analyzed dataset and also a better perception of time variations since it is a best and most reliable method compared to the vector approach.

The functional method allowed detecting ammonium, turbidity and conductivity outliers for several months in 2007, 2008, 2009 and 2010. A plausible explanation for most of them seems to be the occurrence of rainfalls. Rainfalls wash frequently large amounts of insoluble substances into the rivers causing turbidity increases.

There are robust evidences that the outliers obtained by the proposed innovative method are correct. The analysis of the outlier data by using the functional technique provides an estimation of the common dispersion in relation to every concentration level in the studied range (Cuevas and Fraiman, 1997; Peng and Qi, 2008; Barnes and Chu, 2010). To summarize, it can be concluded that this methodology of the functional analysis can be used for the determination of water quality in other rivers with success. Indeed, we have performed the analysis using those specific basic and integrative parameters according to rules on water quality. Thus, the developed method here is used for automatic detection of outliers in rivers but it can also be extended to the analysis of different types of water bodies (i.e., lakes, reservoirs, etc.).

Acknowledgments

This study was possible thanks to the experimental dataset from the Automated Water Quality Information System (AWQIS) located in the basin of the river Miño collected by the Miño-Sil Basin Authority and the Cantabrian Basin Authority (Ministry of Agriculture, Food and Environment of the Government of Spain). At the same time, the authors wish to acknowledge the computational support supplied by the Department of Mathematics at University of Oviedo and “Centro Universitario de la Defensa” at University of Zaragoza. Finally, we would like to thank Anthony Ashworth for his revision of English grammar and spelling of the manuscript.

References


