Imputing missing data using grey system theory and the biplot method to forecast groundwater levels and yields

Jelena Ratković, Dušan Polomčić, Zoran Gligorić and Dragoljub Bajić*

Faculty of Mining and Geology, University of Belgrade, Đušina 7, 11000 Belgrade, Serbia;
(jmocevic45@gmail.com, dusan.polomcic@rgf.bg.ac.rs, zoran.gligoric@rgf.bg.ac.rs, *correspondence: dragoljub.bajic@rgf.bg.ac.rs)


Article history:
Manuscript received January 28, 2020
Revised manuscript accepted February 04, 2022
Available online June 23, 2022

Abstract
Groundwater management is one of today’s important tasks. It has become necessary to seek out increasingly reliable methods to conserve groundwater resources. Dependable forecasting of the amounts of groundwater that can be abstracted in a sustainable manner requires long-term monitoring of the groundwater regime (rate of abstraction and groundwater levels). Monitoring of the groundwater source for the town of Bečej, Serbia had been disrupted for multiple years. The objective of the paper is to assess the possibility of reinterpreting the missing data, or, in other words, to reconstruct the operation of the groundwater source and its effect on groundwater levels. At the Bečej source, groundwater is withdrawn from three water-bearing strata comprised of fine- to coarse-grained sands. Historic data are used to reconstruct the operation of the Bečej source between 1st of October 1980 to 1st of May 2010. The monitored parameters are total source yield and piezometric head at seven observation wells and 14 pumping wells. A data reconstruction methodology was developed, which included the use of an autoregressive (AR) model, a grey model (GM), and the biplot method. The methodology is applied to fill the data gaps during the considered period. The paper also describes the criteria for evaluating the accuracy of the AR model, GM, and biplot method. The proposed data reconstruction approach yielded satisfactory results and the methodology is deemed useful for the Bečej source data, as well as other historic data not necessarily associated with groundwater sources, but also groundwater control and protection systems, as well as hydrometeorological, hydrological and similar uses.

Keywords: groundwater management, groundwater source, forecasting

1. INTRODUCTION
Groundwater is one of the primary sources of water for various uses. The demand for groundwater will steadily rise in the future. This is a growing problem and more and more attention will be devoted to seeking out methods and strategies for progressively efficient groundwater management. In practice, the water supply for a particular area is often provided by developing groundwater sources that deliver sufficient water quantities of a satisfactory quality.

Effective groundwater source management largely depends on local natural conditions, the amount and quality of collected operational data, and the efficient and reliable forecasting of the future water demand and the ability of the hydrogeological setting to meet that demand. Roughly 75% of the water demand in Serbia relies on groundwater. However, there has been no regular monitoring of surface water and groundwater dynamics during a period of economic crisis in the country, which resulted in an underdeveloped observation network. At groundwater sources, regular monitoring was hindered by poor organization and lack of instrumentation and human resources. More recently, the surface water and groundwater observation network has been expanded based on European Union recommendations and the scope and extent of monitoring are now satisfactory.

The need to develop methods that will forecast sustainable rates of groundwater abstraction is becoming increasingly pronounced. Sufficiently reliable prognostic methods would create conditions for effective conservation of groundwater resources, addressing problems such as inefficient use and over-exploitation, which impairs groundwater quality. Reliable forecasts of groundwater levels, groundwater balance, and the velocity and direction of groundwater flow play a major role, in view of the fact that they provide important information about the characteristics of the tapped aquifer. With the exception of groundwater quality, forecasting basically focuses on groundwater levels at a particular rate of abstraction.

Today, many methods are used to forecast groundwater levels, such as Nonlinear Autoregressive Networks (NARX) (GUZMAN et al., 2017; WUNSCH et al., 2018), Artificial Neural Networks (ANN) (NOURIANI et al., 2008; RAKHSHANDEHROO et al., 2012; CHITSAZAN et al., 2015; MOHANTY et al., 2015), Time Series Methods (MIRZAVID & GHAZAVI, 2015), ARIMA and RBF networks (YAN & MA, 2016), Adaptive Neuro-Fuzzy Inference System (ANFIS) (GULDAL & TONGAL, 2009; EMAMGHOLIZADEH et al., 2014), fuzzy neural networks (ALVISI & FRANCHINI, 2011), Fuzzy C-Mean Clustering and Singular Spectrum Analysis (POLOMČIĆ et al., 2017), and hydrodynamic modeling (POLOMČIĆ et al., 2011; POLOMČIĆ et al., 2014).

The objective of the paper is to present a case study of the Bečej groundwater source for drinking water supply (Serbia) and demonstrate the reconstruction of missing data (over a 10-year period), using the autoregressive model, the grey system theory and biplot method, in order to improve decision-making related to groundwater management.

Autoregressive modeling was used as a supplementary method to extract the points in time (dates) of no monitoring. It has been implemented by many other researchers in the fields of...
hydrology (CHAKRABORTY et al, 2009; MEHDIZADEH, 2020), seismology (XIAOJUN et al., 2015), exodynamics (YONGPING et al., 2020), and hydrogeology (RISTIC VAKANJAC et al., 2018).

Many researchers in hydrogeology and similar disciplines have applied the grey system theory to address diverse problems, such as predicting earthquakes (DAOGONG et al., 1999), forecasting of pore pressure changes (MOHAMMED et al., 2010), projecting land subsidence to prevent hazards (TANG et al., 2008), sustainable groundwater management for drinking water supply (JIANFEI et al., 2012), estimating the amount of rainfall and runoff to prevent flooding (YU et al., 2001; KANG et al., 2006), forecasting groundwater levels (MAHMOD et al., 2013; KANG & MAENG, 2016) and predicting and analyzing spring discharges (HAO et al., 2010).

The biplot method is also applied by many researchers in different fields. For example, this statistical method has been very useful in addressing environmental protection issues, to predict the evolution of air pollution (GONZALEZ CABRERA et al., 2006; ZHANG et al., 2018); in mechanics, to solve problems applying SVD (singular value decomposition, which is the basis of the biplot method) (MEIJAARD, 1993; GAI & HU, 2018); in geophysics, to improve the quality of seismic data (MARTINS et al., 2016) and to determine the spatial relationship between deep geologic structures and gold (ZHAO & CHEN, 2010).

Upon reconstruction of the missing piezometric head data and total groundwater source yields, the gaps in the Bečej groundwater source (Bečej GS) data were filled over the study period. Sufficiently reliable input data were provided to assess the feasibility of Bečej GS expansion, predict the long-term effect of groundwater abstraction, and propose optimal solutions for future production wells.

2. STUDY AREA

The study area is the extended area of Bečej GS, Serbia, as shown in Fig. 1. The groundwater source is located west of the city of Bečej. Groundwater from the tapped aquifers is used for both the municipal and industrial water supplies. Between 1980 and 2010, 25 production wells and seven observation wells were drilled at Bečej GS. Fourteen production wells are currently in service (Fig. 1). The others have been decommissioned. The wells tap several aquifers. Their individual discharge capacities are 3 – 16 l/s.

Figure 1. Location map of the study area.
Table 1. Well designations and tapped aquifers.

<table>
<thead>
<tr>
<th>Active wells</th>
<th>Tapped aquifers</th>
</tr>
</thead>
<tbody>
<tr>
<td>B0-1</td>
<td>second, third</td>
</tr>
<tr>
<td>B0-2</td>
<td>first, second</td>
</tr>
<tr>
<td>B1-2/1</td>
<td>second, third</td>
</tr>
<tr>
<td>B1-3/1</td>
<td>first, second, third</td>
</tr>
<tr>
<td>B1-5</td>
<td>second, third</td>
</tr>
<tr>
<td>B1-6</td>
<td>second, third</td>
</tr>
<tr>
<td>BII-1</td>
<td>first, second, third</td>
</tr>
<tr>
<td>BII-2</td>
<td>first, second, third</td>
</tr>
<tr>
<td>BII-3</td>
<td>first, second, third</td>
</tr>
<tr>
<td>BII-4/1</td>
<td>first, second, third</td>
</tr>
<tr>
<td>BII-5</td>
<td>third</td>
</tr>
<tr>
<td>BII-6</td>
<td>second, third</td>
</tr>
<tr>
<td>BS-5</td>
<td>first, second, third</td>
</tr>
<tr>
<td>BS-6</td>
<td>first, second, third</td>
</tr>
</tbody>
</table>

Table 1. includes the designations of the wells and shows the aquifers they capture.

The extended study area is located within the Pannonian Plain and is composed of Precambrian/Palaeozoic, Mesozoic, Miocene, Pliocene and Quaternary rocks. The oldest rocks are crystalline schists and Precambrian/Palaeozoic granites. The youngest rocks are Quaternary in age, with alternating lithological units (gravels and coarse sands underlie sandy silts, silts and clays) (MALEŠEVIĆ, 1984; TERZIN, 1994). Vertically, there are two types of aquifers down to a depth of about 130 m: an unconfined intergranular aquifer in the alluvial and terrace deposits (from the land surface to roughly 30 m – aquifer I) and three confined intergranular aquifers in deeper geological formations (from 60 m to 130 m – aquifers II, III and IV). Bečej GS taps the deeper aquifers of Quaternary age: the first is at a depth of 60 to 80 m, the second at 83 to 100 m, and the third from 110 to 130 m (Fig. 2).

The unconfined aquifer is composed of sands in the upper part and fine-grained, partly clayey sands in the lower part. The bottom of this layer comprises sandy clay. The three deeper aquifers are composed of fine- to medium-grained sands, with some gravel. The sands include a pelitic component, which hinders production/causes well ageing. The aquifers are separated by relatively thin layers of clay, sandy clay and clayey sand. Their thickness ranges from 5 to 20 m. These three aquifers constitute a hydrogeologic complex, which evolved during the Lower and Middle Pleistocene. On a regional scale, they are considered as a single, confined aquifer (MALEŠEVIĆ, 1984; TERZIN, 1994).

Figure 2. A hydrogeological section through the Bečej Groundwater Source. Legend: 1. surface decomposed material; 2. clay; 3. sand; 4. clayey sand; 5. unconfined aquifer; 6. confined aquifer; 7. semi-permeable sediments; 8. confined aquifer head; 9. well; and 10. position of well screen.
3. METHODOLOGY – THE FORECASTING MODEL BUILDING PROCESS

3.1. Problem statement

The study period involving groundwater levels and groundwater abstraction rates at Bečej GS is from 1 October 1980 to 1 May 2010. However, piezometric head was not monitored from 1 February 1989 to 1 June 1998 and total source yield from 1 October 1980 to 1 June 1998. The monitoring network comprised seven observation wells (P1, P2, P3, P4, P5, P6, P7/2), at depths from 120 m to 150 and 14 pumping wells (B0-1, B0-2, BI-2/1, BI-3/1, BI-5, BI-6, BIII-1, BIII-2, BIII-3, BIII-4/1, BIII-5, BIII-6, BS-5, BS-6), at which total groundwater source capacity was gauged (Fig. 3). Based on the data, groundwater levels were observed much more frequently than the total source yield. These two parameters were recorded periodically, but slightly more often in 2006 to assess the water balance and determine groundwater reserves. The present research considers mean monthly values of the parameters, which were the only available data (Fig. 3).

As part of the reported research, a methodology was developed to reconstruct the missing data (10-year period) due to the lack of continuous monitoring. It is shown in Fig. 4 in the form of an algorithm.

In the first step, input data comprising values observed during the course of monitoring were analyzed. The objective of the proposed methodology was to reconstruct the missing data applying the methods indicated in the algorithm. The AR model was used to reconstruct the dates (observation positions) on which piezometric head and source yield were not observed. The accuracy of the AR model was determined on the basis of AIC (Akaike information criterion), according to which the lowest value was assumed to be accurate. After the observation dates were reconstructed, the grey method was applied to arrive at piezometric head values for the period during which there was no monitoring. The input data for the GM were the original and reconstructed dates of observation and the observed piezometric heads. First the GM was tested against the (original) data observed during the study period. If a comparison of the original and reconstructed piezometric head data and MAPE criterion (Table 2) indicated that the accuracy of the model was satisfactory, the approach was acceptable for reconstructing the missing data. If not, the proposed methodology was not suitable for further reconstruction. Given that calculations and the MAPE criterion (Table 2) indicated that the accuracy of GM was satisfactory, it was applied to fill the data gaps. The above methods used to determine the missing observation dates and piezometric heads produced a completed matrix for the biplot method, which was applied to reconstruct the total source yields during the period of no observations. The accuracy of the total source yields derived by the biplot method was determined based on the condition $d < 0.01$ (where:

$$d = \left[ \left( \frac{1}{na} \sum_{i=1}^{na} (x_i - \bar{x}_i)^2 \right)^{1/2} \right], \ y = \left[ \left( \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{n} y_{ij}^2 \right)^{1/2} \right].$$

If the
solution was acceptable, the model was deemed suitable for reconstructing the total source yields. If not, it would have been necessary to look for other, more accurate models. All the calculations were made in Microsoft Excel (AR model, grey theory and biplot method) and MathLab (biplot method). The figures were produced in CorelDraw (Figs. 1, 2, 4) and Microsoft Excel (Figs. 3, 5, 6, 7, 8, 9).

A brief theoretical background of the models used in this research is provided below.

### 3.2. Autoregressive (AR) model
The AR model is commonly applied to process statistical time-series. The AR model was used for predicting missing observation positions. More specifically, it is used to forecast the values of current and future variables, based on the values of the same variable at different times. The variable is described as a function of its values in previous periods \( Z_{t-1}, Z_{t-2} \).

The mathematical expression is (DAI et al., 2015):

\[
X_t = \sum_{i=1}^{p} a_i X_{t-i} + Z_t \tag{1}
\]

where \( p \) is the AR model order and \( Z_t \) is the constant.

In order to construct the AR model, first its order \( (p) \) needs to be determined. The principle is as follows: for the original (observed) data \( (x_1,x_2,...,x_n) \) of order \( p \), the AR model is:

\[
X_t = \sum_{i=1}^{p} a_i X_{t-i} + Z_t \tag{2}
\]

If the original time-series is stationary, the AR process is applied directly to the original dataset. If not, the time-series needs to be transformed into a stationary time-series (sequence). The common transformation approach is first-order differentiation, as was required in the present case.

First-order differentiation is implemented as:

\[
\Delta X_t = X_t - X_{t-1} \tag{3}
\]

Then the AR model can be applied to the differentiated time-series. The ultimate forecasted values are derived from:

\[
\hat{X}_t = X_{t-1} + \Delta \hat{X}_t \tag{4}
\]

where \( \Delta \hat{X}_t \) is the value resulting from the AR process.

The AR model order is defined (determined) applying AIC. AIC is calculated as (CHAKRABORTY et al., 2009):

\[
AIC = \ln \left( \frac{\sum_{i=1}^{N} u_i^2}{N} \right) + \frac{2p}{N} \tag{5}
\]

where \( N \) is the length of the sample, \( p \) is the number of estimated coefficients (\( p \)-order, AR model order) and \( u_i \) are the residuals.

The model where the AIC is the lowest is selected as optimal (most suitable).

Using the AR model, the observation positions (dates) are reconstructed and then used as input data for the GM.

### 3.3. Grey model (GM)

The theory provides an effective solution for the data uncertainty and discretization problem. GMs predict the future values of a time-series, based solely on a set of observed data (recorded, original values in the present case). According to the grey system theory, a general grey model GM(h,N) is the GM in which \( h \) is the order of the differential equation and \( N \) is the number of parameters.

GM(h,N) is defined by the following differential equation (YU et al., 2001):

\[
\sum_{i=0}^{h} a_i \frac{d^{(i)} x_1^{(1)}(k)}{dt^{(i)}} = \sum_{j=2}^{N} b_j x_j^{(1)}(k), k = 1,2,..., n \tag{6}
\]

where \( a_i \) and \( b_j \) are the determined coefficients, \( x_1^{(1)}(k) \) is the sequence of the main factor, \( x_j^{(1)}(k) \) are the sequences of the influential factors, and \( k \) is the time sequence variable.

A single-variable first-order grey model GM(1,1) was used in the present research because there was a single time-series of observations.

The original time-series was represented as (ZHAO-HUI & JING, 2016):

\[
X^{(0)} = [x^{(0)}(t_1), x^{(0)}(t_2),..., x^{(0)}(t_j)], j = 1,..., n, \text{ if:}
\]

\[
\Delta t_j = t_j - t_{j-1} \neq \text{const}
\]
Before proceeding with the GM equation, the accumulating generation operator (AGO) will be defined. AGOs are used in GM to reduce data randomness.

The first-order AGO for \( X(0) \) is defined as (ZHAO-HUI & JING, 2016):

\[
X^{(1)}(t_j) = X^{(1)}(t_{j-1}) + X^{(0)}(t_j) \Delta t_j
\]  

(7)

where \( X^{(1)}(t_j) = \sum_{j=1}^{\infty} X^{(0)}(t_j) \).

The sequence of the generated averages of the neighboring values of sequence \( X^{(0)}(t_j) \) is

\[
Z^{(0)}(t_j) = \left[ z^{(0)}(1), z^{(0)}(2), ..., z^{(0)}(t_j) \right] ,
\]

(8)

where:

\[
z^{(0)}(t_j) = \frac{1}{2} \left[ X^{(1)}(t_j) + X^{(1)}(t_j-1) \right] , \quad t_j = 1, 2, ..., n
\]

(8)

The grey differentiated first-order equation GM(1,1) is (ZHAO-HUI & JING, 2016):

\[
\frac{dx^{(1)}(t)}{dt} = ax^{(1)}(t) + b
\]

(9)

The first term \( x^{(1)}(t_j) \) of the sequence \( x^{(0)}(t_j) (j = 1, 2, ..., n) \) is used as the initial condition of the grey differentiated equation.

The solution to Eq. (9) is a time-dependent function:

\[
x^{(1)}(t) = e^{at} x^{(1)}(1) + a^{-1}(e^{at} - 1)b
\]

(10)

To distinguish between \( a \) and \( b \), Eq. (9) is discretized. In other words, the difference in grey derivative \( I \) of interval \( (t_{j-1}, t_j) \) yields \( x^{(0)}(t_j) = a x^{(1)}(t_j) + b \) and thus:

\[
x^{(0)}(t_j) = a(z^{(1)}(t_j) + z^{(1)}(t_{j-1})) + b \quad (i = 1, 2, ..., n; k = 2, 3, ..., m)
\]

(11)

Coefficients \( a \) and \( b \) are obtained by the least squares method:

\[
a(b)^T = (B^TB)^{-1}B^TY
\]

(12)

where:

\[
B^{(1)}(t_j) = \begin{bmatrix}
\frac{1}{2}(x_1^{(1)}(t_j) + x_1^{(1)}(t_1)) & 1 \\
\frac{1}{2}(x_2^{(1)}(t_j) + x_2^{(1)}(t_2)) & 1 \\
\vdots & \vdots \\
\frac{1}{2}(x_m^{(1)}(t_j) + x_m^{(1)}(t_{m-1})) & 1
\end{bmatrix}
\]

(13)

\[
Y = \begin{bmatrix}
x^{(0)}(t_2), x^{(0)}(t_3), ..., x^{(0)}(t_n)
\end{bmatrix}^T
\]

The time sequence response formula, that is, the solution of (9) is:

\[
x^{(0)}(t_j) = e^{a(t_j-t_1)} x^{(1)}(t_1) + a^{-1}(e^{a(t_j-t_1)} - 1)b, \quad j = 1, 2, ..., n
\]

(14)

The fitting forecast value of \( x(0) \) can be obtained by the inverse AGO, as follows (LIU & LUO, 2010):

\[
x^{(0)}(t_j) = \frac{x^{(1)}(t_j) - x^{(1)}(t_{j-1})}{\Delta t_j}, \quad j = 2, ..., n
\]

(15)

The accuracy of the model is represented by the mean absolute percentage error (MAPE) (ZHAO-HUI & JING, 2016):

\[
MAPE = \frac{1}{n} \sum_{j=1}^{n} \left| \frac{x^{(0)}(t_j) - x^{(0)}(t_j)}{x^{(0)}(t_j)} \right|
\]

(16)

where:

\[
e(t_j) = \frac{x^{(0)}(t_j) - x^{(0)}(t_j)}{x^{(0)}(t_j)}
\]

(17)

In the present case, MAPE was used to assess model accuracy. The model accuracy criterion is shown in Table 2 (WANG & PHAN, 2016).

The GM (1,1) was used to reconstruct piezometric head data during the 10-year period of discontinued observations (from 1 February 1989 to 1 June 1998, Fig. 3). The completed observation positions (dates) and piezometric heads represented input data for the biplot method.

### 3.4. Biplot method

The biplot method involves the imputing of arbitrary values to missing data in order to complete the matrix, and then calculating SVD using only two components (ARCINIEGAS-ALARCON et al., 2014). SVD is the basis of the biplot method. In the reported case study, the biplot method was used to forecast missing data on total source yields from 1 October 1980 to 1 June 1998.

The initial data for constructing the matrix included:

1. Original and reconstructed observation positions (dates) on which groundwater level and total source yield were observed;
2. Piezometric head data – original and reconstructed using the GM (1,1); and
3. Observed source yield data.

The constructed nxp matrix denoted by \( X \) comprised \( x_{ij} \) \((i=1, ..., n; j=1, ..., p)\) elements, where the missing data were symbolized by \( x_{ij}^{aux} \). In the first step, the missing values were imputed in their respective columns, to complete matrix \( X \). In the specific case, observed average discharges (i.e. known values) were used to fill the data gaps.

The completed matrix \( X \) was standardized:

\[
p_{ij} = \frac{(x_{ij} - m_{ij})}{s_j}
\]

(18)

where:

\( m_{ij} \) is the mean deviation of the j-th column,
\( s_j \) is the standard deviation of the j-th column, and
\( p_{ij} \) are the standardized elements.

The matrix with \( p_{ij} \) elements was denoted by \( P \). The next step was to calculate SVD of matrix \( P \). SVD of a standardized matrix is a fundamental and prominent approach in modern numerical linear algebra. The SVD algorithm represents an approximation of the main matrix by smaller matrices (MEI-

<table>
<thead>
<tr>
<th>MAPE (%)</th>
<th>Forecast (prediction) accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1</td>
<td>High</td>
</tr>
<tr>
<td>1-5</td>
<td>Good</td>
</tr>
<tr>
<td>5-10</td>
<td>Acceptable</td>
</tr>
<tr>
<td>&gt;10</td>
<td>Unacceptable (low)</td>
</tr>
</tbody>
</table>
JAARD, 1993). According to this theorem, any matrix $A$ (m×n) can be decomposed into three matrices, as:

$$ A = ULV $$

1. $U$ is the orthogonal m×m matrix,
2. $L$ is the diagonal m×m matrix with positive elements on the main diagonal, referred to as singular values of $A$, and all the other terms are equal to zero.

The singular values are generally sorted in descending order, on the main diagonal $L$.

Taking into account the first two components, the resulting equation (expression) is:

$$ p_{ij} = \frac{(x_{ij} - m_j)}{s_j} = \sum_{k=1}^{2} \lambda_k a_{ik} y_{jk} + e_{ij} \quad (20) $$

- $\lambda_k$ are the singular values,
- $a_{ik}$ are the eigenvectors of the rows,
- $y_{jk}$ are the eigenvectors of the columns, and
- $e_{ij}$ is the error of row $i$ in column $j$.

The new matrix, composed of $p_{ij}^{(2)}$ elements, was denoted by $P^{(2)}$. All the $p_{ij}^{(2)}$ elements in $P^{(2)}$ were re-assigned their original values, $x_{ij}^{(2)} = m_j + s_j p_{ij}^{(2)}$. This resulted in a new matrix $X^{(2)}$. Filling of data gaps was followed by iterations (back to matrix standardization, repeating the entire process) until the specified condition $\left(\frac{d}{y} < 0.01\right)$ was fulfilled and the imputed missing values stabilized. The iterations then continued until the following condition was met (ARCINIE-GAS-ALARCON et al., 2014):

$$ \frac{d}{y} < 0.01 \quad (21) $$

where:

$$ d = \left[\frac{1}{na}\sum_{i=1}^{na}(x_i - x_i^A)^2\right]^{\frac{1}{2}} $$

and

$$ y = \left[\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{p}(y_{ij})^2\right]^{\frac{1}{2}} \quad (23) $$

- $na$ is the total number of values missing in matrix $X$,
- $x_i$ is the value forecasted for the $i$-th row of missing values in the active iteration,
- $x_i^A$ is the value forecasted for the $i$-th row of missing values in the previous iteration,
- $y_{ij}$ are the original values (not missing) in the $i$-th row and $j$-th column, and
- $N$ is the total number of considered values.

4. RESULTS AND DISCUSSION – APPLICATION OF THE METHODOLOGY TO A REAL CASE STUDY

Before the missing data were reconstructed, the hydrogeological system was analyzed based on data collected in the last 30 years of operation to determine the groundwater dynamics.

The unconfined intergranular aquifer (aquifer I) is relatively thin and the values of its hydraulic parameters are low, so the groundwater reserves are modest. This aquifer is recharged by precipitation and via a good hydraulic link with a river. The aquifer is drained by the pumping wells and by migration into the deeper aquifers, given the semi-permeable floor of aquifer I.

Recharge of the confined intergranular aquifers (II, III, and IV) is highly complex because of the large distances involved, so the recharge process can only be assumed based on the geological and hydrogeological features of the terrain and long-term piezometric levels of the groundwater. In view of the fact that the spread of the aquifers is regional, recharge occurs over a large area where their overlying strata are dominated by sands or in zones where the aquifers are directly linked to the land surface or rivers. Infiltration of precipitation into the confined aquifers has not been confirmed or occurs on a small scale. It is assumed that recharge also takes place through groundwater migration from the first, unconfined aquifer, given that the strata overlying aquifer I are composed of sands, clayey sands and sandy clays, which allow partial groundwater infiltration from aquifer I. This infiltration rate is very slow. The confined aquifers are drained solely by the studied pumping wells and neighbouring groundwater sources.

All the observation wells registered identical piezometric head fluctuation trends – lower in the summer and early autumn months due to higher rates of groundwater abstraction. Given the amount of water withdrawn from these aquifers, this decrease indicates that recharge conditions are satisfactory. The total average monthly yield of the groundwater source suggests that the amount of abstracted groundwater has not changed significantly compared to 10 years ago. The recorded piezometric level ranged from a minimum of 65 m above sea level to a maximum of 76 m.a.s.l., while the total source yield was from 68 l/s to 124 l/s.

The missing data reconstruction results are described below. Records of groundwater abstraction rates since the groundwater source was commissioned (a period of 30 years, Fig. 3 provided input data for applying the GM (1,1) and biplot method to Bečej GS. The missing data were reconstructed according to the algorithm shown in Fig. 4, first the observation dates using the AR model, then the groundwater levels resulting from GM (1,1), and finally the total source yield applying the biplot method. The proposed methodology with regard to the period of no observations is described below.

4.1. Imputing of missing observation positions based on the AR model

The AR model was used to forecast observation data (dates) of piezometric head and total source yield. Observation date is the date (day, month and year) of measurement. For AR modeling
purposes, these dates were converted into cumulative days (Fig. 3 shows the dates of original measurement). Each observation date is identified as an observation position in order to define the total number of measured and later reconstructed values (125 measured and 38 reconstructed).

Given that the original time-series was transient (Fig. 5) and in view of the growing trend resulting from Eq. (3), first-order differentiation (Fig. 6) was needed in order to apply the AR model. Time-series that exhibit a trend are referred to as non-stationary.

Equation (1) was applied to the stationary time-series, the values for the AR models of the second, third and fourth orders were calculated, and then the AIC criterion was used to assess the most accurate solution. AIC was applied to second-, third- and fourth-order AR (AR2, AR3 and AR4). The results based on Eq. (5) are shown in Table 3.

The lowest AIC value was adopted (i.e. second-order AR, AR2).

Figure 7 shows the original dates (in red) from 1 October 1980 to 1 February 1989, based on which the missing dates (shown in blue) were reconstructed, along with subsequent original dates (1 June 1998 to 1 May 2010).

Table 4 is a numerical representation of the forecasted observation positions based on second-order AR. The dates shown in red were reconstructed (there were no actual piezometric head and source yield observations).

The MAPE of AR2 was 2.26 %. In view of the large number of missing observation dates, this error was deemed acceptable. Also, AR2 was adopted according to AIC.

4.2. Imputing of missing piezometric head data based on the GM

The input data used in this research were considered to be a grey system, with some known and some unknown information, so that it could be analyzed applying the grey system theory. The first step was to analyze the input data from 1 October 1980 to 1 May 2010, and the second step to apply a GM (1,1). The GM (1,1) was used to reconstruct piezometric head values for the period during which the observation wells (P1, P2, P3, P4, P5, P6, P7-2) were not monitored. The first step was to verify the accuracy of the GM (1,1) using the observed data.

Table 3. AIC values for AR2, AR3 and AR4.

<table>
<thead>
<tr>
<th></th>
<th>AIC(AR2)</th>
<th>AIC(AR3)</th>
<th>AIC(AR4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.43</td>
<td>7.44</td>
<td>7.44</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Reconstructed observation positions based on second-order AR.

<table>
<thead>
<tr>
<th>Observation position</th>
<th>Day</th>
<th>Observation position</th>
<th>Day</th>
<th>Observation position</th>
<th>Day</th>
<th>Observation position</th>
<th>Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>19</td>
<td>1827</td>
<td>37</td>
<td>3315</td>
<td>55</td>
<td>4924</td>
</tr>
<tr>
<td>2</td>
<td>183</td>
<td>20</td>
<td>1888</td>
<td>38</td>
<td>3405</td>
<td>56</td>
<td>5013</td>
</tr>
<tr>
<td>3</td>
<td>274</td>
<td>21</td>
<td>1919</td>
<td>39</td>
<td>3494</td>
<td>57</td>
<td>5102</td>
</tr>
<tr>
<td>4</td>
<td>397</td>
<td>22</td>
<td>2009</td>
<td>40</td>
<td>3583</td>
<td>58</td>
<td>5192</td>
</tr>
<tr>
<td>5</td>
<td>517</td>
<td>23</td>
<td>2100</td>
<td>41</td>
<td>3673</td>
<td>59</td>
<td>5281</td>
</tr>
<tr>
<td>6</td>
<td>639</td>
<td>24</td>
<td>2162</td>
<td>42</td>
<td>3762</td>
<td>60</td>
<td>5371</td>
</tr>
<tr>
<td>7</td>
<td>762</td>
<td>25</td>
<td>2374</td>
<td>43</td>
<td>3851</td>
<td>61</td>
<td>5460</td>
</tr>
<tr>
<td>8</td>
<td>854</td>
<td>26</td>
<td>2435</td>
<td>44</td>
<td>3941</td>
<td>62</td>
<td>5549</td>
</tr>
<tr>
<td>9</td>
<td>882</td>
<td>27</td>
<td>2527</td>
<td>45</td>
<td>4030</td>
<td>63</td>
<td>5639</td>
</tr>
<tr>
<td>10</td>
<td>1004</td>
<td>28</td>
<td>2557</td>
<td>46</td>
<td>4119</td>
<td>64</td>
<td>5728</td>
</tr>
<tr>
<td>11</td>
<td>1127</td>
<td>29</td>
<td>2618</td>
<td>47</td>
<td>4209</td>
<td>65</td>
<td>5817</td>
</tr>
<tr>
<td>12</td>
<td>1248</td>
<td>30</td>
<td>2680</td>
<td>48</td>
<td>4298</td>
<td>66</td>
<td>5907</td>
</tr>
<tr>
<td>13</td>
<td>1309</td>
<td>31</td>
<td>2740</td>
<td>49</td>
<td>4388</td>
<td>67</td>
<td>5996</td>
</tr>
<tr>
<td>14</td>
<td>1370</td>
<td>32</td>
<td>2831</td>
<td>50</td>
<td>4477</td>
<td>68</td>
<td>6085</td>
</tr>
<tr>
<td>15</td>
<td>1462</td>
<td>33</td>
<td>2954</td>
<td>51</td>
<td>4566</td>
<td>69</td>
<td>6175</td>
</tr>
<tr>
<td>16</td>
<td>1613</td>
<td>34</td>
<td>3046</td>
<td>52</td>
<td>4656</td>
<td>70</td>
<td>6264</td>
</tr>
<tr>
<td>17</td>
<td>1674</td>
<td>35</td>
<td>3136</td>
<td>53</td>
<td>4745</td>
<td>71</td>
<td>6353</td>
</tr>
<tr>
<td>18</td>
<td>1735</td>
<td>36</td>
<td>3226</td>
<td>54</td>
<td>4834</td>
<td>72</td>
<td>6443</td>
</tr>
</tbody>
</table>
Equations (7) through (15), described in the theoretical section, were applied in that order to the original time-series of observed piezometric head data. This yielded the predicted piezometric head values. The objective was to check the accuracy and applicability of the GM to the case study or, in other words, to the original measured data (Table 5). Then Eq. (16) was applied to the reconstructed time-series to check the calculation error and Table 2. was used to assess whether the GM was applicable to the original time-series and if it yielded acceptable solutions.

Table 5. shows the observed (original) and reconstructed piezometric heads at observation well P1. It is apparent that the difference between the observed and reconstructed values was acceptable according to the criterion from Table 2 and that the model could be used for the discontinued period from 2 January 1989 to 6 January 1998.

The errors are shown in Table 6. The observed and reconstructed piezometric head values agreed rather well, given the ten-year period of discontinued
groundwater monitoring (Table 5). Based on the criterion shown in Table 2, the conclusion was that the error was acceptable in view of the fact that the accuracy of the reconstructed values was high, so the solution was deemed acceptable.

4.3. Imputing of missing source yield data based on the biplot method

The biplot method was used to reconstruct the missing source yield data, from 1 October 1980 to 1 July 1998. The input data in this case included original and reconstructed:
– observation dates,
– piezometric heads and
– observed source yield data.

The first step, before the biplot method, was to complete matrix $X$ for calculating the missing total source yield values. The averages of the observed source yields were used for the missing values, $x_{ij}^{\text{aux}}$. The completed matrix included the observation dates (original and reconstructed), piezometric heads (original and reconstructed), observed source yields and average source yields (in place of those that needed to be reconstructed). Matrix $X$ was then standardized applying Eq. 18, resulting in standardized matrix $P$. This was followed by the SVD of matrix $P$ (in Mat-Lab), based on Eq. 19. Then Eq. 20 was applied, resulting in matrix $P^{(2)}$. After applying $\hat{x}_{ij} = m_j + s_j p^{(2)}_{ij}$, all the elements of that matrix reverted to original values in matrix $X^{(2)}_{\text{exp}}$. The missing $x_{ij}^{\text{aux}}$ in the original matrix $X$ were completed with the corresponding values of $\hat{x}_{ij}^{(2)}$ from matrix $X^{(2)}$. This was followed by iterations, back to matrix standardization, and the entire process was repeated. Iterations lasted until the condition $\frac{d_y}{y} < 0.01$ was fulfilled (ARCINIEGAS-ALARCON et al., 2014), which occurred in the second iteration (Table 7).

Table 8 and Figure 8 show the observed (original) and reconstructed source yields.

### Table 6. Reconstructed piezometric head errors at all observation wells.

<table>
<thead>
<tr>
<th>Well</th>
<th>Max. abs. error</th>
<th>MAPE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>6.08</td>
<td>1.60</td>
</tr>
<tr>
<td>P2</td>
<td>4.42</td>
<td>1.34</td>
</tr>
<tr>
<td>P3</td>
<td>3.31</td>
<td>1.16</td>
</tr>
<tr>
<td>P4</td>
<td>4.51</td>
<td>1.40</td>
</tr>
<tr>
<td>P5</td>
<td>3.61</td>
<td>1.02</td>
</tr>
<tr>
<td>P6</td>
<td>3.81</td>
<td>1.27</td>
</tr>
<tr>
<td>P7-2</td>
<td>5.83</td>
<td>1.66</td>
</tr>
</tbody>
</table>

### Table 8. Observed (original) and reconstructed source yields – iteration 2.

<table>
<thead>
<tr>
<th>Day</th>
<th>Q Original time-series</th>
<th>Q Reconstr. time-series</th>
<th>Day</th>
<th>Q Original time-series</th>
<th>Q Reconstr. time-series</th>
<th>Day</th>
<th>Q Original time-series</th>
<th>Q Reconstr. time-series</th>
</tr>
</thead>
<tbody>
<tr>
<td>6483</td>
<td>95.31</td>
<td>93.18</td>
<td>8614</td>
<td>106.28</td>
<td>106.08</td>
<td>9832</td>
<td>98.10</td>
<td>98.13</td>
</tr>
<tr>
<td>6514</td>
<td>96.99</td>
<td>95.85</td>
<td>8675</td>
<td>116.34</td>
<td>113.58</td>
<td>9862</td>
<td>99.40</td>
<td>98.39</td>
</tr>
<tr>
<td>6545</td>
<td>104.60</td>
<td>99.85</td>
<td>8737</td>
<td>109.64</td>
<td>109.02</td>
<td>9893</td>
<td>100.72</td>
<td>97.37</td>
</tr>
<tr>
<td>6575</td>
<td>101.23</td>
<td>97.75</td>
<td>8767</td>
<td>105.43</td>
<td>108.55</td>
<td>9923</td>
<td>100.29</td>
<td>93.04</td>
</tr>
<tr>
<td>6606</td>
<td>98.06</td>
<td>97.95</td>
<td>8828</td>
<td>96.55</td>
<td>101.54</td>
<td>9985</td>
<td>100.73</td>
<td>94.78</td>
</tr>
<tr>
<td>6698</td>
<td>101.41</td>
<td>100.84</td>
<td>8949</td>
<td>94.83</td>
<td>97.10</td>
<td>10014</td>
<td>97.35</td>
<td>93.02</td>
</tr>
<tr>
<td>6879</td>
<td>93.87</td>
<td>94.02</td>
<td>8979</td>
<td>96.86</td>
<td>97.34</td>
<td>10045</td>
<td>95.14</td>
<td>92.17</td>
</tr>
<tr>
<td>7063</td>
<td>91.33</td>
<td>86.23</td>
<td>9071</td>
<td>105.12</td>
<td>101.47</td>
<td>10075</td>
<td>95.83</td>
<td>92.19</td>
</tr>
<tr>
<td>7148</td>
<td>99.94</td>
<td>92.97</td>
<td>9102</td>
<td>103.90</td>
<td>100.80</td>
<td>10106</td>
<td>95.86</td>
<td>92.81</td>
</tr>
<tr>
<td>7245</td>
<td>105.14</td>
<td>101.06</td>
<td>9132</td>
<td>103.47</td>
<td>99.98</td>
<td>10136</td>
<td>97.43</td>
<td>94.89</td>
</tr>
<tr>
<td>7276</td>
<td>104.93</td>
<td>102.58</td>
<td>9163</td>
<td>97.93</td>
<td>96.67</td>
<td>10167</td>
<td>97.71</td>
<td>95.77</td>
</tr>
<tr>
<td>7306</td>
<td>98.80</td>
<td>99.41</td>
<td>9193</td>
<td>95.82</td>
<td>94.82</td>
<td>10198</td>
<td>90.37</td>
<td>92.84</td>
</tr>
<tr>
<td>7337</td>
<td>100.77</td>
<td>100.84</td>
<td>9255</td>
<td>97.66</td>
<td>94.33</td>
<td>10228</td>
<td>92.88</td>
<td>92.99</td>
</tr>
<tr>
<td>7367</td>
<td>101.13</td>
<td>101.91</td>
<td>9283</td>
<td>95.60</td>
<td>92.87</td>
<td>10259</td>
<td>87.41</td>
<td>89.06</td>
</tr>
<tr>
<td>7398</td>
<td>100.97</td>
<td>100.85</td>
<td>9314</td>
<td>94.68</td>
<td>91.59</td>
<td>10289</td>
<td>86.34</td>
<td>88.36</td>
</tr>
<tr>
<td>7549</td>
<td>104.38</td>
<td>105.04</td>
<td>9344</td>
<td>96.76</td>
<td>91.21</td>
<td>10320</td>
<td>86.97</td>
<td>88.42</td>
</tr>
<tr>
<td>7610</td>
<td>102.22</td>
<td>103.67</td>
<td>9375</td>
<td>99.93</td>
<td>94.22</td>
<td>10351</td>
<td>83.21</td>
<td>86.20</td>
</tr>
<tr>
<td>7763</td>
<td>100.82</td>
<td>102.24</td>
<td>9436</td>
<td>97.11</td>
<td>94.61</td>
<td>10379</td>
<td>78.58</td>
<td>82.91</td>
</tr>
<tr>
<td>7822</td>
<td>98.87</td>
<td>100.19</td>
<td>9467</td>
<td>93.47</td>
<td>92.56</td>
<td>10410</td>
<td>79.41</td>
<td>83.37</td>
</tr>
<tr>
<td>7853</td>
<td>99.52</td>
<td>101.09</td>
<td>9497</td>
<td>96.28</td>
<td>93.79</td>
<td>10440</td>
<td>77.86</td>
<td>82.76</td>
</tr>
<tr>
<td>7914</td>
<td>99.53</td>
<td>101.06</td>
<td>9528</td>
<td>96.59</td>
<td>92.61</td>
<td>10471</td>
<td>81.66</td>
<td>87.75</td>
</tr>
<tr>
<td>7975</td>
<td>96.76</td>
<td>100.87</td>
<td>9558</td>
<td>92.51</td>
<td>89.59</td>
<td>10501</td>
<td>86.99</td>
<td>90.80</td>
</tr>
<tr>
<td>8067</td>
<td>98.54</td>
<td>103.96</td>
<td>9589</td>
<td>92.61</td>
<td>86.49</td>
<td>10532</td>
<td>82.75</td>
<td>90.24</td>
</tr>
<tr>
<td>8097</td>
<td>96.35</td>
<td>103.06</td>
<td>9620</td>
<td>92.85</td>
<td>86.91</td>
<td>10563</td>
<td>78.07</td>
<td>87.35</td>
</tr>
<tr>
<td>8159</td>
<td>94.78</td>
<td>100.60</td>
<td>9648</td>
<td>94.74</td>
<td>91.53</td>
<td>10593</td>
<td>67.28</td>
<td>81.33</td>
</tr>
<tr>
<td>8248</td>
<td>105.43</td>
<td>105.68</td>
<td>9679</td>
<td>95.66</td>
<td>90.18</td>
<td>10654</td>
<td>87.77</td>
<td>91.09</td>
</tr>
<tr>
<td>8309</td>
<td>112.72</td>
<td>112.49</td>
<td>9709</td>
<td>94.59</td>
<td>93.39</td>
<td>10685</td>
<td>87.84</td>
<td>90.26</td>
</tr>
<tr>
<td>8371</td>
<td>123.38</td>
<td>118.85</td>
<td>9740</td>
<td>97.82</td>
<td>95.18</td>
<td>10744</td>
<td>105.61</td>
<td>99.02</td>
</tr>
<tr>
<td>8462</td>
<td>112.00</td>
<td>110.48</td>
<td>9770</td>
<td>103.30</td>
<td>99.40</td>
<td>10775</td>
<td>85.27</td>
<td>86.95</td>
</tr>
<tr>
<td>8524</td>
<td>113.51</td>
<td>111.08</td>
<td>9801</td>
<td>100.54</td>
<td>97.68</td>
<td>10805</td>
<td>81.44</td>
<td>83.75</td>
</tr>
</tbody>
</table>
It is apparent in Figure 8. that the original and reconstructed source yield data agree rather well, over the entire study period (from 1 October 1980 to 1 May 2010).

Table 9. shows the errors in the 2nd iteration.

Figure 9. shows the Bečej GS observed and reconstructed piezometric head and source yield data.
Figure 9. shows that the trends of the piezometric head and source yield data are consistent with those registered in situ. The reason for the lack of fluctuations of the reconstructed values is attributable to the long period of time during which the considered parameters were not observed, as the applied methodology cannot produce more accurate solutions. However, the proposed methodology was highly effective in the study area, given that all the solutions were within the set criteria and the errors were in the predefined range. The models used in this research yielded satisfactory results and it is believed that they can be applied to fill data gaps in the study area (Bečej, Serbia). The developed methodology can also serve as a tool for upgrading other models (e.g., hydrodynamic) in groundwater management, when available input data are insufficient for this type of analysis. The level of error associated with the proposed methodology indicates that it can be applied to other groundwater sources and groundwater management systems.

5. CONCLUSION

An adequate groundwater source monitoring database is essential for efficient decision making in groundwater resources management. When such data are lacking, it is a challenge to provide sufficiently reliable groundwater for forecasting the future status of a groundwater source and the regime of the captured groundwater.

The paper described a methodology developed to reconstruct missing piezometric head and source yield data because monitoring of the Bečej groundwater source (Bečej GS) had been discontinued for a period of ten out of 30 years. The accuracy of the reconstructed values was first assessed in respect of the observed (original) data. MAPE showed that the applied models were highly accurate, indicating that the methodology could be applied to the entire time series of data observed during the study period.

Temporarily discontinued monitoring of piezometric head and source yield is one of the problems that had an adverse effect on the reliability of decisions relating to the feasibility of expanding the groundwater source in the future, in view of the potential of the area and the predicted efficiency of groundwater abstraction over a long period of time.

The proposed methodology effectively solves that problem. The implementation of the proposed algorithm, resulting in complete input data, improved source expansion solutions based on one of many methods for predicting groundwater dynamics. In the specific case, future research will focus on hydrodynamic modeling with reconstructed and filled data gaps for Bečej GS.

The proposed, verified methodology can be applied to groundwater abstraction systems, but also other systems, such as groundwater control (mines, hydraulic structures, agriculture, industrial zones, and urban areas), groundwater protection, and groundwater control (mines, hydraulic structures, agriculture, groundwater abstraction systems, but also other systems, such as modeling with reconstructed and filled data gaps for Bečej GS. The implementation of the proposed algorithm, resulting in complete input data, improved source expansion solutions based on the predefined range. The models used in this research yielded satisfactory results and it is believed that they can be applied to fill data gaps in the study area (Bečej, Serbia). The developed methodology can also serve as a tool for upgrading other models (e.g., hydrodynamic) in groundwater management, when available input data are insufficient for this type of analysis. The level of error associated with the proposed methodology indicates that it can be applied to other groundwater sources and groundwater management systems.

6. ACKNOWLEDGMENT

The authors express their gratitude to the Serbian Ministry of Education, Science and Technological Development for supporting scientific research, which is essential for the advancement of a knowledge based society.

REFERENCES


