Nonparametric tests for stationary analysis in hydrological data

Uilson Ricardo Venâncio Aires, Guilherme Barbosa Reis, Jasmine Alves Campos

Universidade Federal de Viçosa-UFV, Departamento de Engenharia Agrícola, Avenida Peter Henry Rolfs, SN, Viçosa-MG, Brazil. CEP: 36570-000. E-mail: uvaires@gmail.com, guilherme.eaa.reis@gmail.com, jasminealvescampos@gmail.com.

ARTICLE INFO
Received 09 Mar 2019
Accepted 06 Sep 2019
Published 06 Sep 2019

ABSTRACT
One of the main problems for water resources management systems has been the climate change and the intensification of anthropogenic activities in river basins. In this context, this work aimed to analyze the dynamics of land use and cover and its influence on temporal variability on streamflow data. The behavior of hydrological data (streamflow and rainfall) over time was analyzed by applying the nonparametric tests of Mann Kendall and Pettitt. Images derived from orbital sensors using the Random Forest classifier assessed the anthropogenic influence in the area, land use, and cover classification. The rainfall data did not present significant changes over time, according to the applied tests. However, the low annual flow and average annual flow presented nonstationary behavior, with a trend of reduction over time. As rainfall did not change in its patterns over time, the main reason associated with the changes in streamflow regimes was associated with the changes in land use and land cover, especially in the areas for crops, that had an increase of 48% in the study period, which can contribute to increase the demand for water and affect the streamflow. The results obtained confirm the importance of this study for water management systems to adapt itself to the changes in hydrological behavior over time.

Keywords: Land use and land cover, hydrological dataset, Mann Kendall, Pettitt.

Introduction
The climate change and intensification of anthropogenic activities in river basins caused several problems for water management. Mainly because the global temperature elevation has been affecting the rainfall values and evapotranspiration, which impact in regional hydrological processes (Wang, He & Dong, 2019).

On the other hand, the effects of human activities in hydrological data can be divided into direct effects and indirect effects; the direct effects are related to the use of water for irrigation, industrial and agricultural activities, which directly obtained water from rivers (Milly et al., 2008). Indirect effects include afforestation and urban construction which change the land use and land cover patterns and, consequently, the hydrological cycle, affecting the spatial and temporal distribution of water resources (Diem, Hill & Milligan, 2018).

Many studies have been conducted to understand the effects of changes in vegetal cover in rivers streamflow. For the low annual flow and annual average flow, it has a consensus that the deforestation in the short period tends to increase the water available in the river in that streamflow regime. It occurs because there is a reduction in the consumption of water by plants, and for this reason, a decrease in evapotranspiration values (Andréassian, 2004; Bosch & Hewlett, 1982; Farley, Jobbagy & Jackson, 2005; Molina et al., 2012). The suppression of vegetation caused different effects on the maximum annual flow: increased, decreased, and even null effects (Andréassian, 2004).

The analysis of the hydrological behavior over time often uses the Mann Kendall, and Pettitt's nonparametric methods, because of its simplicity and robustness (Ding, Wang & Lu, 2018; Ebrahimian et al., 2018; Xue et al., 2018; Zhang et al., 2018). The Pettitt test, besides verifying stationary in data series, also allows identifying the period in which the change occurred (Ebrahimian et al., 2018).

The analysis of change in the streamflow over time is fundamental for planning and
management of water resources since the low water availability can jeopardize the socio-economic development of the whole river basin. In this context, the study aimed to evaluate the dynamics of land use and cover and its influence on the temporal variability of streamflow.

**Material and Methods**

*Location of the study area*

The study was carried out in the drainage area (2.7554 km²) of the streamflow gauge station Belo Vale (code: 4071000) located in the region upstream of the Paraopeba river basin (BHRP) (Figure 1). This basin is located in the Unidade de Planejamento e Gestão dos Recursos Hídricos (UPGRH) Alto São Francisco (SF03). The basin has a drainage area of 13,642 km² and contains 45 municipalities, being of extreme importance for the urban water supply of Belo Horizonte.

![Figure 1. Drainage area location of the streamflow gauge station used in this study in the Paraopeba river basin. Source: Aires, Reis & Campos (2019).](image)

*Hydrological dataset*

The hydrological dataset used were obtained from the National Water Agency (Agencia Nacional de Aguas, ANA), where the streamflow data has 50 years of observed data from 1968 to 2017. Also were used three rain gauge stations with observed data from 1942 to 2017, with 50 years of observed data each.

Due to the need of use longer and updated historical series, non-consistent data (level 1) of the streamflow gauge and rain gauge stations were used, it means that the data from the period from 2005 to 2017 are those collected in the field, and do not pass through consistent analyses. However, a careful analysis of all the data was made in order to detect errors that could affect the hydrological analyzes.

The need to use more extended historical series refers to the fact that, in shorter series, natural fluctuations in hydrological data can be due to nonstationary behavior (WMO, 1988).

*Stationary analysis of streamflow and rain dataset*

The nonparametric tests of Mann Kendall
and Pettitt were applied to verify the behavior of hydrological data over time, considering the level of significance of 5% (Mudbhatkal et al., 2017). The tests were performed using software R, for the Mann Kendall and Pettitt test, the Kendall and Trend package were used, respectively.

In the Mann Kendall test, the variable S, for a series of n data, is calculated from the sum of the sgn signals of the difference, in pairs, of all values of the series xi in relation to the subsequent values of the data series xj, expressed in Equation 1 (Salviano et al., 2016).

\[
S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{sgn}(x_j - x_i) \quad \text{Eq.(1)}
\]

when \( n \geq 10 \), the variable S can be compared with a normal distribution, and the variance \( \text{Var}(S) \) can be expressed by Equation 2.

\[
\text{Var}(S) = \frac{(n-1)(2n+5)-\sum_{i=1}^{n} t_i(2i+5)}{18} \quad \text{(Eq. (2))}
\]

where \( t_i = \) number of repetitions of an extension i.

The trend observed in the historical series is considered to be increasing or decreasing according to the positive and negative values of the \( Z_{MK} \) index, respectively. According to the S signal, the \( Z_{MK} \) index is calculated from Equation 3.

\[
Z_{MK} = \frac{S - 1}{\sqrt{\text{Var}(S)}}; \text{ for } S > 0
\]

\[
Z_{MK} = 0; \text{ for } S = 0 \quad \text{Eq. (3)}
\]

\[
Z_{MK} = \frac{S + 1}{\sqrt{\text{Var}(S)}}; \text{ for } S < 0
\]

The hypothesis \( H_0 \) is rejected, that is, the series presents a temporal trend, if the absolute value of \( Z_{MK} \) is higher than the tabulated value \( Z_{\alpha/2} \), or the P-value is lower than the level of significance adopted.

The Pettitt test, also used to verify the stationary of the data, allows, in the case of series that do not present stationery, to identify the period in years of the change occurred. This test is a version of the Mann-Whitney statistic \( U_t \), \( N \), applied to verify if two samples \( x_1, ..., x_i \), and \( x_{i+1}, x_n \) belong to the same population, the test statistic \( U_t \), \( N \) obtained through Equation 4 (Uliana et al., 2015).

\[
U_{t,N} = U_{t,N} + \sum_{j=1}^{N} \text{sgn}(x_i - x_j) \quad \text{para } t = 2,...,N
\]

The absolute maximum value of \( |U_t|, Ni \) indicates the position of the possible change in the behavior of the data series. Thus, the statistic \( k(t) \) (Equation 5) represents the change point \( t \) associated with the level of significance \( p \) (Equation 6).

\[
k(t) = \max_{1 \leq t \leq N} |U_{t,N}|
\]

\[
\rho \equiv 2 \exp \left\{ -6 \left( \frac{k^2}{N^2 + N^2} \right) \right\} \quad \text{Eq.(6)}
\]

where \( p \) = critical value associated with the level of significance and \( N \) = number of samples in the hydrological data. If \( p \) is less than the level of significance adopted, the null hypothesis (\( H_0 \) = no significant changes in the data series) is rejected.

The trend analysis was applied in order to study the streamflow and rainfall behavior over time. For this, the hypothesis of stationary was verified, considering hydrological year for the dataset of average annual streamflow, and the annual peak and low streamflow. For the rain dataset, the tests were applied for the total annual rainfall, in each of the three rain gauge stations with influence in the study area.

**Obtaining and processing images**

Orbital images derived from the Thematic Mapper (TM)/Landsat 5 and Operational Land Imager (OLI)/Landsat 8 were used to perform the land use and land cover classification.

The acquisition and processing of those orbital images were performed on the Google Earth Engine (GEE) cloud computational platform. The script used to obtain the orbital images was adapted from the Google Earth Engine API (GEE, 2017), which presents several examples of scripts for manipulation of data on the GEE.

In the script used, it is possible to highlight two functions of great assistance to obtain the orbital images, mainly due to the presence of clouds. Through the CFmask band, available in the images of Landsat Surface Reflectance (SR), it is possible to detect and remove the clouds (Function 1).

\[
\text{var maskClouds = function(image)}
\]

\[
\{ \text{var cs = cloud_shadows(image)}; \quad \text{Func.(1)}
\]

\[
\text{var c = clouds(image)};
\]

Aires, U.R.V.; Reis, G.B.; Campos, J.A.
image = image.updateMask(cs);
return image.updateMask(c)}

However, this process resulted in places where there is no information in the image (NoData); for this reason, a second function was used to fill those gaps (Function 2).

This procedure creates a single image of a period of interest. For example, if there are four orbital images from January to March, Function 1 will remove all the clouds presents in those images. On the other hand, Function 2 will create an average image after the first procedure.

This procedure allows filling the gaps because as the images are from different periods, there is a possibility that the clouds were removed in different parts of the images.

In this study, an average image was created to performed the land use and land cover classification, using the period of 01/1985 to 12/1986 from the TM/LandSat 5 and for 01/2017 to 12/2017 from the OLI/LandSat 8.

Classification of land use and land cover

The classification of land use and land cover was done using the Random Forest classifier with the support of software R, using the package labgeo.

It was collected training samples in polygons shape containing 9 to 12 pixels of the Landsat images with the support of ArcGis® Desktop 10.1 software. The number of training samples and the classes of land uses used in this study can be found in Table 1.

Table 1. Classes of land use identified in the study area and the number of samples.

<table>
<thead>
<tr>
<th>Land uses</th>
<th>Number of samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forest</td>
<td>50</td>
</tr>
<tr>
<td>Cultivation area</td>
<td>40</td>
</tr>
<tr>
<td>Pasture</td>
<td>50</td>
</tr>
<tr>
<td>Urban area</td>
<td>20</td>
</tr>
<tr>
<td>Water resources</td>
<td>20</td>
</tr>
<tr>
<td>Bare soil</td>
<td>20</td>
</tr>
</tbody>
</table>

The cultivation areas comprise eucalyptus forest and annual crops because they had a similar behavior of their texture when we were collecting the samples, thus that land-use class can be confused by the classifier. Mining areas seem to represent most of the bare soil areas.

It was used 75% of the total sampled areas for training, and 25% for validation of the Random Forest classifier and the metric for the evaluation of the accuracy of the classifier used was the Kappa (K) index, given by Equation 7.

$$
k = \frac{n \sum_{i=1}^{y} x_{i} - \sum_{i=1}^{y} (x_{i} x_{i})}{n^2 - \sum_{i=1}^{y} (x_{i} x_{i})}
$$

where $K =$ estimate of the Kappa coefficient, dimensionless, $x_{i}$ is the value of row i and column i; $x_{ii}$ is the sum of line i and $x_{+i}$ is the sum of column i of the confusion matrix; n represents the total number of samples representing the land use observed, and y the total number of land use classes.

The interpretation of the values obtained for $K$ was performed based on the classification presented in Table 2, proposed by Landis and Koch (1977).

Table 2. Interpretation of $K$ values.

<table>
<thead>
<tr>
<th>$K$</th>
<th>Level of agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0.00</td>
<td>None</td>
</tr>
<tr>
<td>0.00 – 0.20</td>
<td>Minimal</td>
</tr>
<tr>
<td>0.20 – 0.40</td>
<td>Weak</td>
</tr>
<tr>
<td>0.40 – 0.60</td>
<td>Moderate</td>
</tr>
<tr>
<td>0.60 – 0.80</td>
<td>Strong</td>
</tr>
<tr>
<td>0.80 – 1.00</td>
<td>Almost perfect</td>
</tr>
</tbody>
</table>


The Land Change Modeler (LCM) tool of the Idrisi Selva® software was used to verify the changes in land use and land cover between the two analyzed periods. This tool has been used for estimating changes in land use as a result of building projects, such as highways, reservoirs, deforestation, among other anthropogenic actions (Abuelaish & Olmedo, 2016).

Results

Change analysis of land use and land cover in the study area

Figure 2 shows the land use and cover for the period of 1985/86 (Figure 2a) and 2017 (Figure 2b) in the study area.

The Kappa index obtained for the 1985/86 classification was 0.91 and the one for the 2017 period, 0.98, considered excellent according to the ranges in Table 3. Although the high value got with the kappa index for both land use classification images, it was possible to verify some errors, mainly between bare soil and urban areas. Thus, a higher number of samples could improve the classification because it would enable them to have more parameters for training and validation of the
random forest classifier.

The investigation of changes in land use and land cover using the LMC can be observed in Figure 3. Figure 3(a) is the loss of area in km² and the Figure 3(b) is the percentage of changes in land use and land cover from one class to another between 1985/86 to 2017.

Figure 2. Land use and land cover classification: a) period corresponding to 1985/86; b) period corresponding to 2017. Source: Aires, Reis & Campos (2019).

Figure 3. Changes in land-use and land cover between 1985/86 and 2017: a) gains and loss of area for each land-use class in km²; b) Percentage of land use and land cover changes of each class from a previous period.
to the later period. Source: Aires, Reis & Campos (2019).

Figure 3a highlights that the higher gain in the land use classes in the period 1985/86 to 2017 was in the cultivated areas, with an increase of 554 km², demonstrating an intensification in agricultural activities in the area. On the other hand, the areas with pasture and forest were the ones that suffered the most reductions in the analysis period, with 615 km² and 315 km², respectively. Forest areas have decreased mainly in the central region of the study area.

In Figure 3b, the area of water resources increased by 85% in its area in the period of 1985/86. This increase was motivated mainly because of the construction of a dam in the region near the mining area. The cultivation areas also showed an increase in the previous period, with 48%. On the other hand, the bare soil was the one that presented the most reduction, with 44%, motivated mainly by the regeneration of the vegetation in the areas that was occupied by mining activities.

**Stationary Analysis of streamflow and rainfall dataset**

Figure 4 shows the annual values of the peak (a), low (b) and average (c) streamflow rates for the period from 1968 to 2017.

![Graph of streamflow](image)

Figure 4. Values of the peak (a), low (b) and average (c) streamflow between 1969 and 2017. Source: Aires, Reis & Campos (2019).

Figure 4 highlights the reduction trend of the average and low streamflow, while the peak streamflow shows a natural variation, despite a high value observed in 1995.

Figure 5 shows the annual rainfall behavior of the three rain gauge stations with the highest influence area in the downstream region of the Paraopeba river.
Figure 5. Annual rainfall behavior in the three rain gauge stations used in this study, from 1942 to 2017. a. Rain gauge station code 02043005; b. Rain gauge station code 02043013; c. Rain gauge station code 02044007. Source: Aires, Reis & Campos (2019).

In Figure 5, the rainfall behavior appears to have a natural fluctuation, with a decrease in the rain volume in recent years. To confirm this, the application of nonparametric tests is necessary.

Before using the Mann Kendall test, it is necessary to verify the correlation of the hydrologic dataset. If the values are highly correlated, it is necessary to apply the Mann Kendall test in conjunction with block bootstrapping to verify the differences in partial sets. Figure 6 shows the correlation analysis done for the streamflow data. On the other hand, the Pettitt test can be applied directly to the dataset.

Figure 6 shows that few values are higher than the dashed line, considered the acceptable correlation limit without having to perform the second method of analysis. Therefore, the Mann Kendall test can be applied directly to the streamflow dataset.

Figure 7 presents the total and partial correlation analysis of the rainfall dataset used in this study.

Figure 7 shows that the values of the total and partial correlations are also within an acceptable limit for the application of the Mann Kendall test. Thus, it is possible to apply this test directly to the rainfall dataset of the study area.

The results of the Mann Kendall and Pettitt tests, considering the significance level of 5%, can be verified in Table 3.

Table 3. The results of the Mann Kendall and Pettitt test application on the streamflow and rainfall dataset of the study area.

<table>
<thead>
<tr>
<th>Streamflow regime</th>
<th>Mann Kendall</th>
<th>Pettitt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tau of Kendall</td>
<td>P-value</td>
</tr>
<tr>
<td>Peak</td>
<td>0.0939</td>
<td>0.3403</td>
</tr>
<tr>
<td>Low</td>
<td>-0.351</td>
<td>0.0003</td>
</tr>
<tr>
<td>Average</td>
<td>-0.159</td>
<td>0.1046</td>
</tr>
</tbody>
</table>
It can be observed in Table 3 that in the Mann Kendall test, a trend (P-value <0.05) reduction (tau of Kendall <0) was observed only for the low streamflow. On the other hand, the Pettitt test indicated nonstationary behavior for both low and average streamflow.

It would be essential to apply the third test to confirm the results observed with the Pettitt test since the tests indicate different results. For the rainfall data, the tests confirm the observations drawn in Figure 5: annual rainfall volume had natural variation over time in the study area. This results can be indicative that the trends observed in the low and average streamflow may link to the anthropogenic activities in the study area.

![Graphs of total and partial correlation of streamflow dataset](image)

Figure 6. Total and Partial correlation of the streamflow dataset for the study area. Source Aires, Reis & Campos (2019).
Figure 7. Total and partial correlation of the rainfall dataset for the study area. Source: Aires, Reis & Campos (2019).

Figure 8 shows the period in which the changes occurred in the streamflow dataset, through the application of the Pettitt test.
Figure 8. Periods in which the modifications occurred in the streamflow dataset identified with the application of the Pettitt test. Source: Aires, Reis & Campos (2019).

Figure 8 displays a change in the average streamflow patterns occurred near 2000, where the data slightly exceeded the critical value.

Figure 8 also points out that the low streamflow presented two changing points, the former occurring close to 1990, with a trend to increase the values of this streamflow regime, the latter is a trend of decreasing near 2010.

This result may be related to land use, and land cover in the study area, which may have reduced water consumption at critical times, such as irrigation or suppression of vegetation, which reduces evapotranspiration and withdrawal of soil water by plants roots (Farley et al., 2005; Pruski, 2009). Moreover, in the later period, there may an intensification of water use, as well as an increase in the degradation of the area, making it difficult to infiltrate the water in the soil and reducing the recharge of the soil water (Pruski, 2009).

Discussion

The results indicate the importance of hydrological behavior study for management and planning of water resources. The low streamflow, for example, it is very important to calculate how much water users can take from river, fundamental to avoid conflicts of water users in river basins (Almeida & Curi, 2016).

It was observed a trend reduction of this streamflow regime; the water availability is also decreasing over time. Thus, a water resource management model based on the concept that streamflow behavior does not change along the time it is not appropriate. Such water resource management could affect water availability and prejudice the quality of the aquatic environment (Silva et al., 2015).

The increase in the cultivation areas might have influenced the reduction of low and average streamflow observed. Crops have a higher Leaf Area Index than grass, consequently, a higher evapotranspiration value and water demand. Several studies associate the reduction in the streamflow regimes to the increase in vegetal cover
(Andréassian, 2004; Bosch & Hewlett, 1982; Farley et al., 2005; Molina et al., 2012).

This behavior is explained by the fact that reforestation generates a substantial increase in evapotranspiration (ET) (Calder, 1986), besides the higher capacity of withdrawal of the water of the ground due to the root’s depth (Engel et al., 2005; Farley et al., 2005; Zhang et al., 2001).

However it is essential to highlight that the relationship pointed out by the authors refer to specific local conditions, because soil characteristics, such as: texture, organic matter content, hydraulic conductivity, porosity (Jiang et al., 2018; Pinheiro et al., 2009), and even the conditions of the vegetation cover previous to the modifications in the area (Jiang et al., 2017) may influence the observed relationship.

Conclusion

The two streamflow regimes (average and low streamflow) studied displayed a nonstationary behavior, with a trend of reduction over time. This observation confirms the importance of studies for the adequacy of water resources management systems to these modifications, mainly to avoid water use conflicts;

The anthropogenic activities in the drainage area of the streamflow gauge station used in this study, especially the cultivation areas, influenced the reduction of the streamflow, once there may have been an increase in the water demand for crops.

Acknowledgments

The authors would like to express their gratitude to CNPq for the financial support for the project under which this research was carried out.

This study was financed, in part, by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References


Jiang, X. J.; Liu, W.; Wu, J.; Wang, P.; Liu, C.;


