Physical characteristics of pipes as indicators of structural state for decision-making considerations in sewer asset management

Las características físicas de las tuberías como indicadores de la condición estructural y su utilización para tomar decisiones en la gestión de activos del sistema de alcantarillado

Liliana López-Kleine¹, Nathalie Hernández², and Andrés Torres³

ABSTRACT

Sewer deterioration is a problem that affects many cities of the world. This affects the structural state of the sewer systems, as well as its hydraulic capacity and the service level. As a consequence, the sewer system stakeholders are working on the development of a proactive sewer management to make decision in time and avoid public emergencies. Therefore, the objective of this work was to predict the variable state using a clustering algorithm (k-means) in Bogotá’s sewer pipes based on its physical characteristics. Among the most representative results was to find a relationship between pipes’ characteristics and their structural state (chi-squared). Furthermore, the slope and ground level variables were the most related ones to the state of the pipes. The detected relationships are linear and can be used to make management decisions when pipes are clustered and the clusters are mapped on a principal component plane.

Keywords: k-means, sewer asset management, cluster analysis, principal components analysis (PCA), proactive sewer management, sewer pipes, structural pipes state, Bogota’s sewer system.

RESUMEN

El deterioro de los sistemas de alcantarillado es un problema que afecta a las ciudades, no solo en su estado estructural sino también en su capacidad hidráulica y nivel de servicio. En consecuencia, los encargados del sistema de alcantarillado están trabajando en el desarrollo de una gestión proactiva para tomar decisiones a tiempo y evitar emergencias públicas. Es por esto que el objetivo de este trabajo fue predecir la condición de las tuberías en la ciudad de Bogotá utilizando algoritmos tipo cluster (k –means), para discriminar las tuberías que tienen buena condición estructural de las que no. Entre los resultados más sobresalientes se encontró una relación entre las características estructurales de las tuberías y su estado (prueba Chi – cuadrado) siendo la pendiente y la profundidad las variables más relacionadas con el estado de las tuberías. Adicionalmente, estas relaciones encontradas resultaron lineales al agrupar las tuberías en un plano de componentes principales.

Palabras clave: k-means, gestión de sistemas de alcantarillado, cluster, análisis de componentes principales (ACP), gestión proactiva de alcantarillados, tuberías de alcantarillado, condición estructural de tuberías de alcantarillado, sistema de alcantarillado de Bogotá.

Received: April 23rd 2016
Accepted: October 12th 2016

Introduction

As a consequence of the growth of cities, urban water systems are exposed to increased pressures in terms of climate change, environmental pollution, limited resources and aging infrastructure (Ferguson et al., 2013). Drainage systems, which present alarming aging and deterioration rates, are part of the cities’ infrastructure developed over several years (Osman, 2012). As a consequence of their structural deterioration, most of the sewer systems are being every time more prone to fail (Ward & Savic, 2012). This impacts directly on the level of service and quality of life in the communities (Micevski et al., 2002; Osman 2012, Liu & Kleiner, 2013).
Multiple factors influence deterioration of pipes such as their physical characteristics (diameter, length, depth, material, type of joints), installation processes, external factors (characteristic of the supporting soil, soil usage, environmental characteristics) and other factors such as age, type of pipe and inappropriate upkeeping (Davis et al., 2001). More recently, factors such as climate change, soil change and demographic increase have been reported as influencing pipes deterioration (Kleidorfer et al., 2013).

Although, in other countries, several models for planning the maintenance of sewer systems exist (Saegrov, 2006; Mashford et al., 2010), most of them are based on complete and appropriate information, which is not available for the Colombian case. Information on sewer systems inspections is sparse (coverage is low) and the quality of the information is not guaranteed (Rodríguez et al., 2012). For example, the coverage of inspection per year in Bogotá’s sewer system is estimated to be 2%, meaning that the average time between two inspections is 50 years, which is very low compared to international standards (Alluche & Freure, 2002, U.S. EPA, 1999).

Traditionally, in Colombia water service companies have taken on charge asset maintenance and operation with a reactive approach of upkeeping (solving the problem after failure). Nevertheless, experiences show that this approach can be more expensive than using a proactive one (Rodríguez et al., 2012).

Given that sewer pipes are close to completing their useful life cycle, it is foreseeable that in the following years management of infrastructures will be prioritized over the development of new ones. For now it is very important that variables indicating the state of pipes are measured. Therefore, it is crucial that statistical methods predicting in any way the state of pipes in Bogotá, based on measurable characteristics, are developed. These methods need to take into account the percentage of inspected sewer networks, the frequency and the quality of the inspection.

In this study we investigate if a prediction of the variable state is possible using a clustering algorithm. To illustrate our findings the original variable state and the new constructed clusters are mapped on a Principal Component Space. The above-mentioned results allow a first approach on estimating approximately the structural state of sewer pipes and discriminating those with a good state (state 1) from those which need revision (state 5). These results can therefore be used for decision making regarding planning detailed inspections, maintenance, replacements and overall public expenses.

**Materials and Methods**

**Data**

For evaluating the structural condition of pipes of the sewer system in Bogotá, a database obtained using CCTV between 2007 and 2011 was made available by the public aqueduct and sewer systems company in Bogotá, Empresa de Acueducto y Alcantarillado de Bogotá (EAAB) (Figure 1). This database contains information about the physical characteristics of pipes, their location and structural state. The structural state was obtained applying the norm NS-058 (EAAB, 2001) on 3563 waste and rain sewer pipes. Figure 1 shows the location of the inspected sewer pipes from 2007 to 2011 (black lines) and the whole sewer network of Bogotá (gray lines).

The characteristics that were retained due to their possible relationship with the structural state of pipes were: (i) slope, (ii) diameter, (iii) type of material, (iv) age, (v) ground level at the beginning of the pipe, (vi) ground level at the end of the pipe, (vii) depth at the beginning of the pipe, (viii) depth at the end of the pipe, (ix) surface type at ground level, (x) type of pipe and other factors such as the geographical coordinates (east, west). For further analysis only numerical variables were used. The variable state of the pipe indicates the amount of structural damage of a pipe and it was used as an auxiliary variable with the five original categories, but also with three and two categories obtained grouping states two, three, four; and two, three, four, five as shown in Table 1.

![Figure 1. Map of sewer system of Bogotá.](image)
Table 1. Auxiliary variable state of the pipe with five (5), three (3) and two (2) categories.

<table>
<thead>
<tr>
<th>State (5) Original categories</th>
<th>State (3)</th>
<th>State (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

**Statistical Analysis**

A descriptive analysis of the data was undertaken using boxplots in order to identify outliers. Then, the linear correlations between variables were estimated using Pearson’s linear coefficient. These estimates were used to choose which variables were suitable for the Principal Component Analysis because this analysis is based on linear relationships.

The Principal Component Analysis (PCA) was used to resume the structure of the data using linear combinations of the original variables (Lebart et al., 1995). These linear combinations are called Principal Components (PCs), and are obtained by solving an eigenvalue problem which assures that the first PC retains maximum variance of the data (Lebart et al., 1995) and allows a representation of the original data on a lower dimension space.

The clustering algorithm k-means (Hartigan & Wong, 1979) was used to group pipes in a desired number of clusters aiming to retrieve the categories of the variable state. Results were mapped on the PC space in order to observe the obtained behaviors.

The concordance between constructed clusters and original categories of the variable state was evaluated using a chi-square hypothesis test, in which the null hypothesis is no association between variables. Therefore, if the test is rejected, an association between variables is concluded.

**Results and discussion**

Boxplots were constructed for all variables (supplementary material). They allowed detecting an important number of outliers for the ground level variable, but not for the other variables. Therefore, no pipes were eliminated from the data set.

Taking into account that the aim here was to find similarity patterns between pipes across all numerical variables, the linear correlation structure was studied (Table 2). The observed linear relationships are relatively high indicating that a linear multivariate analysis is suitable for these data. This correlation coefficient is very close to one between the variables ground level1 (ground level upstream the pipe section) and ground level2 (ground level downstream the pipe section) indicating that the information contained in them is redundant and therefore we decided to eliminate ground level2. Moreover, the linear relationship of the linear coordinates (X and Y) to all other variables is very weak (lower than rP = 0.50), indicating that there is no linear relationship between numerical variables and location. The variable depth2 also has a relatively low linear relationship to the other ones (maximum is 0.56) (see Table 2).

According with the above results, the authors carried out the PCA on three different scenarios taking into account: (i) All variables of the dataset; (ii) without X and Y variables; and (iii) without localization variables (X and Y) and depth2 variable. The PCs retain in decreasing order as much variance as possible and they are linear combinations (projections) of the original variables. PCA therefore allows resuming a multivariate table in a two-dimension plot in order to observe data structure. Furthermore, we used this plot to show clusters (Figures 4 and 5).

The PCA results on each scenario indicate that for the first scenario the first two PCs retain 47% of the total variance; in the second 57%; and the third 60.5%. For more details on these variance percentages please refer to the supplementary material. The final PCs components that will be used from here for illustration are the ones retaining 60.5% of the total variance with 37.4% on the first PC and 23.1% on the second PC (this result is represented in all figures).

The correlation circle is the projection of the variables on the first two PCs: the first PC (PC1) is represented on the horizontal axis, and the second one (PC2) on the vertical axis. The orthogonal projection of the corresponding vector of each variable over each PC represents the degree of explanation that variable has over each PC. Being the first PC (PC1) the one that explains the most variability of the problem, the variables with a high magnitude of their projection in PC1 will be the ones that most explain the variability of the problem. The correlation circle shown in Figure 2 indicates that the first PC (PC1) is highly explained.
by ground level and slope variables, which means that these variables contribute with the highest amount of information for the construction of this PC. It is important to clear that age shows a small projection with the first two PCs; the small magnitude indicates that age does not explain the variability between pipes as strong as the other variables do it.

Taking into account that the main objective here is to find a relationship between the numerical variables (now resumed on the PC1 and PC2 scatterplot – Figure 3), with the variable state, we mapped this auxiliary variable on the plot indicating which pipes have each of the categories of the variable state (Figure 4). This scatterplot does not show a clear structuration of the structural state variable when it is evaluated with five categories (or structural degrees) because no separation between structural categories (from 1 to 5) in the PC plane is obtained: all the ellipses representing each structural category are overlapped in this PC plane (see Figure 4). Therefore, these categories were reduced to two and three (see Table 1), but no structuring is observed on the scatterplots (for plots see supplementary material). This does not mean that there is no structuring, but at least it is not observable on the PCA scatterplot.

Figure 2. Correlation circle of the final PCA. 60.5% of variance is retained on the first two first PCs.

Figure 3. Scatterplot of pipes on the first two PCs of the final PCA. 60.5% of variance is retained on these two first PCs.

Figure 4. Scatterplot of pipes on the first two PCs, using the variables state with five categories as an auxiliary variable to be mapped.

We mapped k-means constructed clusters on the scatterplot as well (Figure 5), hoping for these clusters emulate the categories of the variable state. It is possible to observe that a structure of clear separation (less overlaps between cluster ellipses) is possible along PC1 (horizontal axis), and therefore mainly explained by ground level and slope variables according to Figure 2 (Figure 5).
Figure 5. Scatterplot of pipes on the first two PCs, using k-means constructed clusters as an auxiliary variable to be mapped. Map of: (a) five clusters, (b) three clusters and (c) 2 clusters.

Furthermore, we investigated if a relationship between the obtained k-means clusters and the variable state existed applying a chi-square test. The null hypothesis of this test is of no association between variables, as explained before. Therefore, a rejection indicates that an association between clusters and original categories exists. Thus, in case of rejection (significant P-values smaller than alpha = 0.05), the obtained clusters are retrieving groups of pipes related to the state and a prediction of the state is made possible. P-values obtained were significant for all three numbers of clusters compared to the original variable state (Tables 3, 4, 5): p-value = $2.2 \times 10^{-16}$ for five clusters, p-value = $0.03165$ for three clusters, and p-value = $0.02726$ for two clusters. This leads to the conclusion that any number of clusters can be used to retrieve the pipe state.

Nevertheless, just knowing that clusters are significantly related to state of the pipe does not inform about 1) which cluster corresponds to which state and 2) the quality of prediction. In order to answer the first question, constructed clusters were mapped on the PC space and compared to the mapping of the variable state. For the second question, contingency tables were constructed (Tables 3, 4, 5) and used to compute the percentage of predictions.

When the frequencies of pipes at each one of the clusters are compared to the categories of the variable state for the case of five categories (Table 3), it can be observed that the highest frequencies are obtained for states 1 and 5. Pipes with state 1 are observed at the left side on the PC space (Figure 5a). They have the highest values for the ground level and slope numerical variables as can be observed on the correlation circle. For the grouping with five clusters and five categories, the cluster grouping more pipes with state 1 is cluster 3: 34%, (197/566 = 0.34). Observing the mapping of five clusters on the PC space (Figure 5a), it is also possible to see that cluster 3 is the one with the center most at the left. For the case of five clusters, clusters 1 and 2 have centers that are very close.

In the same sense, when the grouping with three and two categories/clusters is observed (Table 4 and 5, respectively), frequencies are highest for state 1 for the clusters found left on the PC space: clusters 3 and 2 (Figure 5b and Figure 5c, respectively) are grouping pipes with state 1: 207/602 = 0.34 and 229/66 = 0.35. These results indicate that pipes found on the left of the PC space, with highest values of slope and ground level, can be clustered together in a group containing approximately 34% of pipes with state 1, based only on the numerical variables. Similarly, it is also possible to retrieve pipes with state 5 through the clusters. The clusters that group mainly pipes with this state are cluster 3 (399/1184 = 0.34) for the group of three clusters, and 1 (2027/2902 = 0.70) for two clusters.

Given the results for state 5 with two clusters, in which prediction is 70%, we suggest building two clusters. The cluster grouping pipes with high values of ground level and slope variables would be the one grouping pipes of state 1 (with ca. 34% of pipes of state 1), and opposite to this one on the first PC plane would be the cluster grouping mostly pipes of state 5 (with ca. 70% of pipes of state 5). These pipes belonging to the cluster of pipes with state 5 should be revised in priority. These results indicate that, even though it is not possible to directly predict the structural state from physical characteristics, a relationship exists and therefore models based on them can be proposed. Additionally, this analysis showed that some characteristics are more related than others, such as ground level and slope. In previous studies other variables that showed low relationship to state have been found to influence state of the pipes. These variables are age, diameter and depth (Davies et al., 2001; Saegrov, 2006; Niño et al., 2012). Nevertheless, multivariate analyses are stronger because they allow detecting a global relationship and influence of several factors (Hao et al., 2012).

Particularly in the city of Bogotá and especially for the analyzed database, it has been detected that pipes with high slopes and in elevated neighborhoods (east mountains and Suba), seem to be in better structural conditions than those near the Bogotá river (low slopes and low elevation).
Questions arise on the choice of slopes. Not only hydraulic or topographical conditions should be taken into account, because low slopes could favor hydraulic retention times and increase \( H_2S \) production favoring corrosion of concrete pipes (Jiang et al., 2015). On the other hand, it is possible that pipes near the river could be exposed to higher phreatic levels during rainy seasons depending on soil type and permeability. These infiltrations could cause liquefaction of soils surrounding pipes and therefore loss of supporting material, which is important for the dissipation of strengths (Barragán & Prado, 2014): the direct support of the strengths on pipes could cause fissures and cracks.

Nowadays, the sewer asset management in Bogota is driven in a reactive way (acting after the failure) inducing major risk of collapses in the whole sewer system and spending more money than to develop a proactive asset management plan (Rodriguez et al., 2012). Therefore, these preliminary results should be taken into account in the development of plans focused on proactive sewer asset management with particular characteristics (for example, topography and financial issues) typical of Latin-American cities such as Bogota.

### Conclusions

A relationship between structural characteristics of pipes and their state has been found using a descriptive PCA coupled with k-means clustering. The importance of the different variables has been established, being slope and ground level the most related ones to the state of the pipes. The detected relationships are linear and can be used to make management decisions when pipes are clustered and the clusters are mapped on a principal component plane. Therefore, the statistical approach is pertinent for characterization. Nevertheless, an exact prediction is not possible and further samples, but also other statistical methods based on non-linear data structure (López-Kleine & Torres, 2014), or a combination between statistical methods and learning machine methods such as Fuzzy k-means (Soto & Jiménez, 2011) could be investigated.

### Acknowledgements

The authors would like to thank EAAB for supplying the database information used in this research.

### References


