OUTLIER DETECTION OF LAND SURFACE TEMPERATURE: KÜÇÜKÇEKMECE REGION

Kuşak, L., 1* Küçükali, U. F. 2

1Istanbul Aydin University, Architecture and Design Faculty, Department of Interior Architecture, Istanbul, Turkey (lutfyekkusak@aydin.edu.tr); ORCID ID 0000-0002-7265-245X

2Istanbul Aydin University, Architecture and Design Faculty, Department of Architecture, Istanbul, Turkey (ufkucukali@aydin.edu.tr); ORCID ID 0000-0002-2715-7046

ABSTRACT: Unplanned and rapid urbanization is one of the reasons for the rising surface temperature in urban areas. There is a large amount of literature demonstrating the association of urbanization with surface temperatures. Küçükçekmece Lake, an important lake that has been meeting the utility water needs of Istanbul, and the unplanned and rapid urbanization around it, has been affected by this inevitable change for years. Although surface temperatures generally correlate strongly with each other, very high and very low temperature values should not be disregarded and need to be investigated. The current study was conducted with the assumption that these values could be outliers; thus, they were analyzed using the box plot method for the selected region. Correlations between land surface temperature (LST) values obtained for Küçükçekmece and its vicinity were examined using Landsat Operational Land Imager (OLI) images from June 20, 2016 and June 23, 2017, and LST outliers and regions with common outliers on both days were determined. In the study, 310 LST outliers were identified for June 20, 2016 and 34 LST outliers for June 23, 2017; in both images, 33 outliers were found to be common and these clustered in two different buildings. The reasons for the outliers outside the standard surface temperature values as well as the recommended solutions were discussed.

Keywords: LST, Outlier, Box Plot, Urbanization, Land use/cover
1. INTRODUCTION

According to the data for January 2018 from the Worldometers website (URL 1), 53.25% of the world’s population, about 7.5 billion people, lives in cities. Rapid population growth in cities, which is caused by the fact that the human population abandons rural areas and migrates to urban areas, results in changes in land use as well as unplanned or poorly planned urbanization (Deng et al., 2009), leading to many ecological (Zhao et al., 2008; Wang et al., 2008), climatic (Liu et al., 2017), sociological (Vlahov and Galea, 2002), and environmental problems (Liao et al., 2017). Changes in land use/cover and the resultant problems have been analyzed using ecological (Nacef et al., 2016), economic, and sociological data, and remote sensing images (Shen et al., 2016).

One of the biggest problems brought about by rapid urbanization and consequent changes in land use/cover are changes in surface temperatures in intensely urbanized areas (Weng, 2001; Jenerette et al., 2007; Cui and Shi, 2012; Tayyebi et al., 2018). Numerous studies have been conducted to observe these changes and determine whether there is a relationship between land use/cover and the change in land surface temperatures (Chen et al., 2006; Hasaniou and Mostofi, 2015; Alhawiti and Mitsova, 2016; Li et al., 2017).

The development in remote sensing technology allow land surface temperatures (LST) to be obtained using various satellite and airborne sensors that support the acquisition of thermal infrared information (Kaya et al., 2012; Chen et al., 2013; Gunawardena et al., 2017; Ranagalage et al., 2017).

Atmospheric conditions, geographical factors, and urbanization affect surface temperature values. However, as analyzed in this study, the LST values are affected by unusual local changes and phenomena, such as very low or very high values, which may be called outliers.

Outliers, which are often used in the field of statistics, are widely used in the field of machine learning owing to developments in software and hardware technology that allow storage and processing of much more data. Outliers, also called abnormalities, discordants, deviants, or anomalies, are used in a wide range of activities, including data security, medical diagnosis, various fields of physics, network security, credit card and insurance fraud detection, critical systems, defense activities, and earth science (Iglewicz and Hoaglin, 1993; Bramer, 2007; Kumar and Mathur 2014, URL 2, URL 3). Statistical or model-based models, proximity-based models, linear regression models (Principal Component Analysis (PCA), Least Mean Squares (LMS)), information theory models, high-dimensional outlier-detection methods, and many other methods were developed for the detection of outliers, with data type and data size (univariate/multivariate) being very important in the selection of the method to be used (Sen, 2006). Statistical or model-based models, including standard deviation (SD), Z-scores, and box plots, are preferred for the detection of outliers for univariate data (Olson and Delen, 2008; Han et al., 2011).

A box plot is a graphical method for displaying variation in a group of numerical data. In the literature, box plot is widely used in environmental studies (Xie et al., 1999; Zhang et al., 2009), urban studies (Cipolla and Maglione, 2014), and other scientific studies concerned with outlier detection.

The outlier obtained in some studies may contain highly valuable information (Rousseeuw and Hubert, 2017). As non-spatial data such as temperature, pressure, and image pixel color intensity are obtained on the basis of spatial locations, unusual local changes in these data can be seen as outliers (Aggarwal, 2015).

In this study, the LST values obtained with Landsat 8 Operational Land Imager (OLI) satellite images were prepared given that unusual local changes might affect them. As the obtained LST values are univariate, the box plot method (Tukey, 1997) is used in this study. In addition, box plot was preferred here because it is less sensitive to extreme values (Mendenhall and Sincich 2016).

The interquartile range (IQR) calculations used to construct the box plot were created, and the outlier values obtained from the analysis were discussed and evaluated at the end of the study by taking into account the spatial characteristics (anthropogenic effect, slope, aspect, and water cooling island (WCI)) (Du et al., 2016).

2. DATA AND METHODS

The study comprises three main stages, including acquisition of Normalized Difference Vegetation Index (NDVI), LST images and values of indices; determination of LST values causing outliers by the box plot method; and discussion of potential reasons that might be responsible for LST values described as outliers.

2.1 Study Area

An area within a distance of six km from the Küçükçekmece Lake and its coastline was chosen as the study area. The Küçükçekmece District is situated on the shores of the Marmara Sea on the European side of Istanbul (Fig. 1). The study area is located between 40°57′30″ N–41°06′30″ N and 28°39′0″ E–28°50′30″ E coordinates.

As stated in the 1/100.000 environmental plan of Istanbul, in accordance with east-west development, the study area includes Küçükçekmece Lake, residential areas, industrial areas, agricultural areas, sports facilities, educational facilities, and airport functions. As is the case in Istanbul, the study area has faced highly intensive urbanization for the last 30 years (Kucukali and Kışak, 2017).
\[ \alpha = 1.438 \times 10^{-3} \text{mK}, \text{ and } \epsilon \text{ is the surface emissivity.} \]

There is a relationship between land cover and emissivity. The use of NDVI has often been the preferred method for demonstrating this relationship (Van de Griend and Owe 1993; Valor and Caselles, 1996). The NDVI-based threshold method was then developed for NDVI-based studies (Sobrino et al., 2001; Sobrino et al., 2004). In this method, three main threshold ranges consisting of soil, vegetation, and mixed areas were determined. Certain emissivity values are used for soil and vegetation, but they are created with the emissivity formula calculated using NDVI values for mixed areas. Currently, the three main threshold ranges defined by Sobrino (2004) have increased and different emissivity values and calculation methods for these ranges are being developed (Stathopoulou et al., 2007; Stathopoulou and Cartalis, 2007; Xie et al., 2012; Tang et al., 2015).

In this study, the method adopted by Shen, which is used in the LST calculation, is applied. Here, emissivity values of 0.9923, 0.923, and 0.986 for water areas, urban impervious and bare soil areas, and vegetation areas, respectively, are accepted. The formula in Eq. (2) for mixed areas is also used (Shen et al., 2016).

\[ \epsilon = 1.0094 + 0.047 \ln(\text{NDVI}) \]  

Obtaining the pixel values for the LST images is another step of the study to detect outliers. The tools in ARCGIS 10.2 were exploited to obtain pixel values. With the help of spatial analysis tools used concurrently, multipoint values were obtained from the clipped images. LST, NDVI, Normalized Difference Built Index (NDBI), slope (URL 4), aspect (URL 5) pixel values, and the distance of their spatial location to the lake were obtained using a spatial analysis tool for both dates. During interpretation of the values obtained, correlations between data were investigated first by forming a correlation matrix.

A box plot was used to determine the pixel values of the obtained values, which show temperature values contrary to the data set, and to discuss the situations that could be caused by these outliers. MS Excel and RStudio software are preferred for calculating outliers. IQR is used in the box plot method developed by Tukey. The interquartile range is the range between the first and third quartiles (the edges of the box). Tukey took into account any data point that fell outside either 1.5 times the IQR below the first or 1.5 times the IQR above the third quartile to be “outside” or “far out” (Tukey, 1977).

3. RESULTS

NDVI is an index often used in the detection of live vegetation in the area. NDVI values of the study area from two different days were evaluated. It is clear that the NDVI values were close to 1 in and around the northern parts of the lake in both images (Fig. 2). NDVI is also used intensively to calculate LST values.

\[ \alpha = 1.438 \times 10^{-3} \text{mK}, \text{ and } \epsilon \text{ is the surface emissivity.} \]

There is a relationship between land cover and emissivity. The use of NDVI has often been the preferred method for demonstrating this relationship (Van de Griend and Owe 1993; Valor and Caselles, 1996). The NDVI-based threshold method was then developed for NDVI-based studies (Sobrino et al., 2001; Sobrino et al., 2004). In this method, three main threshold ranges consisting of soil, vegetation, and mixed areas were determined. Certain emissivity values are used for soil and vegetation, but they are created with the emissivity formula calculated using NDVI values for mixed areas. Currently, the three main threshold ranges defined by Sobrino (2004) have increased and different emissivity values and calculation methods for these ranges are being developed (Stathopoulou et al., 2007; Stathopoulou and Cartalis, 2007; Xie et al., 2012; Tang et al., 2015).

In this study, the method adopted by Shen, which is used in the LST calculation, is applied. Here, emissivity values of 0.9923, 0.923, and 0.986 for water areas, urban impervious and bare soil areas, and vegetation areas, respectively, are accepted. The formula in Eq. (2) for mixed areas is also used (Shen et al., 2016).

\[ \epsilon = 1.0094 + 0.047 \ln(\text{NDVI}) \]  

Obtaining the pixel values for the LST images is another step of the study to detect outliers. The tools in ARCGIS 10.2 were exploited to obtain pixel values. With the help of spatial analysis tools used concurrently, multipoint values were obtained from the clipped images. LST, NDVI, Normalized Difference Built Index (NDBI), slope (URL 4), aspect (URL 5) pixel values, and the distance of their spatial location to the lake were obtained using a spatial analysis tool for both dates. During interpretation of the values obtained, correlations between data were investigated first by forming a correlation matrix.

A box plot was used to determine the pixel values of the obtained values, which show temperature values contrary to the data set, and to discuss the situations that could be caused by these outliers. MS Excel and RStudio software are preferred for calculating outliers. IQR is used in the box plot method developed by Tukey. The interquartile range is the range between the first and third quartiles (the edges of the box). Tukey took into account any data point that fell outside either 1.5 times the IQR below the first or 1.5 times the IQR above the third quartile to be “outside” or “far out” (Tukey, 1977).

3. RESULTS

NDVI is an index often used in the detection of live vegetation in the area. NDVI values of the study area from two different days were evaluated. It is clear that the NDVI values were close to 1 in and around the northern parts of the lake in both images (Fig. 2). NDVI is also used intensively to calculate LST values.

\[ \alpha = 1.438 \times 10^{-3} \text{mK}, \text{ and } \epsilon \text{ is the surface emissivity.} \]

There is a relationship between land cover and emissivity. The use of NDVI has often been the preferred method for demonstrating this relationship (Van de Griend and Owe 1993; Valor and Caselles, 1996). The NDVI-based threshold method was then developed for NDVI-based studies (Sobrino et al., 2001; Sobrino et al., 2004). In this method, three main threshold ranges consisting of soil, vegetation, and mixed areas were determined. Certain emissivity values are used for soil and vegetation, but they are created with the emissivity formula calculated using NDVI values for mixed areas. Currently, the three main threshold ranges defined by Sobrino (2004) have increased and different emissivity values and calculation methods for these ranges are being developed (Stathopoulou et al., 2007; Stathopoulou and Cartalis, 2007; Xie et al., 2012; Tang et al., 2015).

In this study, the method adopted by Shen, which is used in the LST calculation, is applied. Here, emissivity values of 0.9923, 0.923, and 0.986 for water areas, urban impervious and bare soil areas, and vegetation areas, respectively, are accepted. The formula in Eq. (2) for mixed areas is also used (Shen et al., 2016).

\[ \epsilon = 1.0094 + 0.047 \ln(\text{NDVI}) \]  

Obtaining the pixel values for the LST images is another step of the study to detect outliers. The tools in ARCGIS 10.2 were exploited to obtain pixel values. With the help of spatial analysis tools used concurrently, multipoint values were obtained from the clipped images. LST, NDVI, Normalized Difference Built Index (NDBI), slope (URL 4), aspect (URL 5) pixel values, and the distance of their spatial location to the lake were obtained using a spatial analysis tool for both dates. During interpretation of the values obtained, correlations between data were investigated first by forming a correlation matrix.

A box plot was used to determine the pixel values of the obtained values, which show temperature values contrary to the data set, and to discuss the situations that could be caused by these outliers. MS Excel and RStudio software are preferred for calculating outliers. IQR is used in the box plot method developed by Tukey. The interquartile range is the range between the first and third quartiles (the edges of the box). Tukey took into account any data point that fell outside either 1.5 times the IQR below the first or 1.5 times the IQR above the third quartile to be “outside” or “far out” (Tukey, 1977).
LST images generated using Eq. (1) and Eq. (2) as well as other emissivity values (Fig. 3) showed that the surface temperature in close vicinity to the lake was lower than those of the inner parts of the study area, as expected. Similarly, the surface temperature of the northern areas, where the NDVI value was close to 1, was also low. On the other hand, the surface temperature was noted to be high in the eastern and western parts of the lake.

Following visual interpretation, 256592 values were obtained by the ARCGIS spatial analyst tool to determine the numeric counterparts of surface temperatures and LST average, SD, Min., Max., and R2 values of the two different LST images obtained (Table 1). Table 1 demonstrates that there is an average drop of 0.82 °C in the selected region in the LST image for June 23, 2017. The LST correlation for June 20, 2016 was 0.965, and that for June 23, 2017 was 0.988.

Table 1  LST values and correlations

<table>
<thead>
<tr>
<th>Year</th>
<th>Average</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>6/20/16</td>
<td>36.73</td>
<td>3.27</td>
<td>25.68</td>
<td>50.28</td>
<td>0.965</td>
</tr>
<tr>
<td>6/23/17</td>
<td>35.91</td>
<td>3.93</td>
<td>23.24</td>
<td>52.05</td>
<td>0.988</td>
</tr>
</tbody>
</table>

The correlations between LST, NDVI, NDBI, slope, aspect pixel values, and distance of their spatial location to the lake were investigated. As expected, the highest negative correlation was found between NDVI and LST. In addition, there was no significant relationship between the other parameters. Moreover, the correlation between LST values of the two different years was 0.889.

Despite this high correlation, very low and high LST values were found when the LST values were listed in increasing order. A box plot analysis (Fig. 4) was performed to determine whether these values were outlier values.

According to the results, the upper level LST value was 46.72 °C, while the lower level LST value was 27.12 °C in 2016. For the year 2017, the upper level LST value was found to be 48.42 °C, while the lower level LST value was found to be 23.66 °C.

In 2016, 234 items of data were below the lower level, while 84 were above the upper level. A total of 318 outlier data were found. One datum of 2017 was below the lower level, whereas 33 were above the upper level. A total of 34 data were determined to be outliers.

After the outlier points determined for the year 2016 were excluded, the correlation value increased to 0.9675, while the correlation value for the year 2017 increased to 0.9883. LST correlations for both days were unchanged and remained at 0.889, compared to the correlations before exclusion of the outlier data.

However, when outlier data are the focus (Figs. 5 and 6), this may lead to results that are too important to neglect, especially in micro-scale studies, although there is not any noticeable change.

When 318 outlier points determined by the box plot analysis of the LST image from June 20, 2016 were examined, 234 points at the lower level were general distributed along the lake shoreline, seaside, and wetlands, whereas 84 outlier points were in the Atatürk Airport area, located in the southeastern part of the image.
When 34 outlier points detected by the box plot analysis of the LST image from June 23, 2017 were examined, one point at the lower level was at the seaside, whereas 33 outlier points were densely distributed in the Atatürk Airport area, located in the southwestern part of the image. When the points where the overlapping outlier points determined by the box plot analysis of the LST images from June 20, 2016 and June 23, 2017 were examined, it is noteworthy that the points at the lower level did not actually overlap, whereas 33 points at the upper level did overlap. When position checks of these points were made, they corresponded to the CNR EXPO (first building) and Atatürk Airport Warehouse (second building); the reasons for this were then investigated (Fig. 7).

Figure. 6 2017 LST outliers images

According to the results obtained by the box plot method of LST data from 2016, outlier temperature values lower than 27.12 °C and higher than 46.72 °C were ascertained. Visual assessment of these outlier data revealed that the outlier values lower than 27.12 °C in particular were located on the lake shoreline and seaside. This result was ascribed to similar atmospheric conditions, similar geomorphologic properties of the locations of these outliers, and the cooling effects of the water surface. On the other hand, the outliers higher than 46.72 °C were densely clustered in the southeastern region of the study area (Fig. 7). These outliers were located in the CNR EXPO (first building) and Atatürk Airport Warehouse (second building), which are in the region of the Atatürk Airport.

Additionally, on the basis of the obtained by the box plot method of LST data in 2017, outlier temperature values higher than 48.42 °C were determined. They completely matched the locations of the 2016 LST outliers, which were the CNR EXPO (first building) and the Atatürk Airport Depot (second building) buildings. This result suggests that the massive size of these buildings, determined according to the results of both analyses, the technical properties of their roofing material, and the positioning of the building’s air-conditioning elements have an impact on the formation of outliers.

5. CONCLUSION

In this study, the box plot method developed by Tukey was used because the data were univariate, and took into account that unusual local changes might have affected the LST values obtained from Landsat 8 OLI satellite images prepared to include the area within a distance of six km from Küçükçekmece Lake and its vicinity (Tukey 1997). The IQR calculations used to construct the box plot were undertaken and the outlier values obtained from the analysis were discussed and evaluated at the end of the study by taking into account the spatial characteristics (anthropogenic effect, slope, aspect, and WCI).

According to the results obtained by the box plot method of LST data from 2016, outlier temperature values lower than 27.12 °C and higher than 46.72 °C were ascertained. Visual assessment of these outlier data revealed that the outlier values lower than 27.12 °C in particular were located on the lake shoreline and seaside. This result was ascribed to similar atmospheric conditions, similar geomorphologic properties of the locations of these outliers, and the cooling effects of the water surface. On the other hand, the outliers higher than 46.72 °C were densely clustered in the southeastern region of the study area (Fig. 7). These outliers were located in the CNR EXPO (first building) and Atatürk Airport Warehouse (second building), which are in the region of the Atatürk Airport.

Additionally, on the basis of the obtained by the box plot method of LST data in 2017, outlier temperature values higher than 48.42 °C were determined. They completely matched the locations of the 2016 LST outliers, which were the CNR EXPO (first building) and the Atatürk Airport Depot (second building) buildings. This result suggests that the massive size of these buildings, determined according to the results of both analyses, the technical properties of their roofing material, and the positioning of the building’s air-conditioning elements have an impact on the formation of outliers.

5. CONCLUSION

In this type of study, temperature, pressure, wind, and pixel values which can be more significant in the interpretation of the results, obtained from satellite images should not be evaluated independently of spatial location characteristics. Thus, the correlations between slopes, aspect, lake distance, NDVI, NDBI, and LST values of the study area were identified in the preliminary phase of the study. However, except for the high correlation between generally accepted NDVI, NDBI, and LST indices for the selected study area, there was no relationship between slope, aspect, distance to the lake, and LST; therefore, these were excluded from the evaluation, but were included as secondary data in the interpretation of the results.

The structures identified as a result of the examination of the outlier locations determined by box plot analysis of the data obtained from the LST images were found to have a very large mass in terms of their architectural design. In such large flat-roofed structures, the reflection and solar orientation design of the roof, planting of an appropriate selection of species in the vicinity of the building, selection of proper roofing material and color, green roof applications, correct positioning of air-conditioning elements, and restoration and improvement of structures may help to reduce the LST value.

It is recommended that further research studies include continuous observation and control of locations with outliers on a seasonal and annual basis, and taking of precautions, where necessary, by expanding the methodology, scale, and spatial and temporal characteristics of the current study.
REFERENCES


Iglewicz, Boris and David Hoaglin (1993), How to Detect and Handle Outliers. American Society for Quality Control, Milwaukee WI.


