CLUSTERING TIME SERIES OF REPEATED SCAN DATA OF SANDY BEACHES

R. Lindenbergh\(^1\)*, S. van der Kleij\(^1\), M. Kuschnerus\(^1\), S. Vos\(^2\), S. de Vries\(^2\),

\(^1\) Dept. Geoscience & Remote Sensing,
\(^2\) Dept. of Hydraulic Engineering,
Delft University of Technology, Delft, The Netherlands

(r.c.lindenbergh, m.kuschnerus, SierdeVries, S.E.Vos)@tudelft.nl, svanderkleij@student.tudelft.nl

Commission II, WG II/10

KEY WORDS: Permanent laser scanning, environmental monitoring, time series, change detection

ABSTRACT:

Sandy beaches are highly dynamic areas affected by different natural and anthropogenic effects. Large changes, caused by a storm for example, are in general well-understood and easy to measure. Most times, only small changes, at the centimeter scale, are occurring, but these changes accumulate significantly over periods from weeks to months. Laser scanning is a suitable technique to measure such small signals, as it is able to obtain dense 3D terrain data at centimeter level in a time span of minutes. In this work we consider two repeated laser scan data sets of two different beaches in The Netherlands. The first data set is from around the year 2000 and consists of six consecutive yearly airborne laser scan data sets of a beach on Texel. The second data set is from 2017 and consists of 30 consecutive daily terrestrial scans of a beach near The Hague. So far, little work has been done on time series analysis of repeated scan data. To obtain a first grouping of morphologic processes, we propose to use a simple un-supervised clustering approach, k-means clustering, on de-leveled, cumulative point-wise time series. The results for both regions of interest, obtained using k=5 and k=10 clusters, indicate that such clustering gives a meaningful decomposition of the morphological laser scan data into clusters that exhibit similar change patterns. At the same time, we realize that the chosen approach is just a first step in a wide open topic of clustering spatially correlated long time series of morphological laser scan data as are now obtained by permanent laser scanning.

1. INTRODUCTION

The surface geometry of our infrastructure and environment is continuously changing due to different physical processes. Monitoring such surface change is important in many cases as either our safety might be affected or because the underlying physical processes are poorly understood. At Dutch sandy beaches both motivations come together. The Dutch beaches with their associated dune landscape directly protect the low lying Dutch hinterland. At the same time, processes affecting the morphology of these continuously changing beaches are partly but not fully understood, (De Schipper et al., 2016).

In recent years, laser scanning has emerged as a well-known technique to closely inspect different landscapes and associated geohazards, (Eitel et al., 2016). Laser scanning combines Light Detection and Ranging (LIDAR) with a mechanism to point a laser ranger in many desired directions, and, in most cases, a method to estimate position and orientation of the scanning system, (Vosselman, Maas, 2010). Laser scanning systems acquire dense point clouds sampling the topography of e.g. a beach environment in a period ranging from minutes to hours. The quality of individual xyz observations is typical at the centimeter level. Laser scanning is relatively expensive, when compared to photogrammetric point cloud acquisition, but requires less or no ground control points and has less difficulties with low texture areas or mapping of areas covered by vegetation.

A variety of sensor systems is available to map geomorphology from air, (Höfle, Rutzinger, 2011), using helicopters, planes or LIDAR drones, from vehicles like cars, (Bitenc et al., 2011), boats and quads, or even from backpack. In addition there are static laser scanners with ranges of up to a few km. These different sensor solutions have in common that all are able to deliver highly detailed point clouds, and after some processing, digital terrain models with a quality that reaches cm level, (Vosselman, Maas, 2010).

Similarly, there have been numerous reports, (Eitel et al., 2016), on detecting changes from repeated scan data. What many reports have in common, is that change methodology either considers only two epochs, that is, before and after a certain event, or consist of methodology that somehow combines consecutive pairwise changes, (Lindenbergh, Pietrzyk, 2015). Current scanning technology makes it now possible to effectively scan the same area over and over again, (Kromer et al., 2017). At several locations, static laser scanners were installed in a permanent way and collected a large number of repeated scans, (Eitel et al., 2016). In (Williams et al., 2018), permanent laser scanning data is used to analyze the frequency of certain type of rock-fall events. One of two data sets considered in this research consists currently of a total number of 5000 hourly scans. Notably at sandy beach environments, where erosion at a certain location is followed by sedimentation in an irregular pattern, pairwise comparisons are not enough to grasp morphological processes.

Next thing to do after pairwise comparison is local trend analysis, (Lindenbergh, Hanssen, 2003, Lindenbergh et al., 2011). Fitting a linear trend is easy and effective in describing change, but, again, two or even a few more parameters may not be sufficient to describe complicated morphological behavior. Therefore here it is considered not to describe changes in multi-epoch data by a limited number of parameters, but rather to cluster...
point-wise time series based on their relative similarity. In this work, clustering is considered on two different data sets. The first data set consists of annual airborne laser scan data of a beach and dune area on the South-West of the Dutch island of Texel, (Lindenberg, Hanssen, 2003) for the years 1998 - 2003. This data set has before been processed using low parameter pre-selected kinetic models. The other data set, (Vos et al., 2017), consist of hourly permanent laser scan data, i.e. repeated point clouds from a terrestrial scanner mounted on a fixed location on top of a hotel, facing the beach. In this work a subset of 30 daily scans is considered. An alternative approach to process data from the same permanent laser scanner is presented in (Anders et al., 2019).

2. METHODOLOGY

In this chapter general methodology is presented that is used to process the two case study data sets considered in this work. Sometimes processing varies. In the following, PLS-KIJKDUIN, refers to the daily scan data from the permanent laser scanning setup at Kijkduin, The Hague, while ALS-TEXEL refers to the yearly airborne laser scan data from Texel.

As input to the proposed work-flow a spatial data set of aligned topographic time series is expected. That is, at each location

\[ p_i = p_i(x_i, y_i), \quad i = 1, \ldots, n, \]

a time series

\[ T_i = (h_i(t_1), \ldots, h_i(t_T)) \]

covering \( T \) epochs is expected. To obtain such input, in some cases a pre-processing step is required to interpolate input data to the same grid. In the current version, no no-data gaps in the time series are allowed. Goal of the method is to cluster time series that show the same behavior through time.

2.1 Point cloud gridding

Data sets in this project had different spatial support in different epochs. To obtain time series at fixed locations, therefore a gridding step was used. Two different approaches were used. In case of PLS-KIJKDUIN, data was interpolated to a regular 1m grid using nearest neighbor interpolation, provided that that neighbor was located within the grid cell. In case of ALS-TEXEL, points were interpolated to a regular 1m grid, using a spline interpolation. Only those grid cells were considered in the time series analysis whose cell center was within 1m of the nearest observation in 2D.

After the gridding step, it is considered if time series at a given grid point are complete. If one or more epochs are missing, a grid point and its associated time series are removed from further processing. In case of PLS-KIJKDUIN, one day was very foggy, resulting in many missing grid points. Therefore this complete day was removed from the analysis.

2.2 De-leveling

Before the actual clustering, first each time series is de-leveled. That is, the mean elevation \( \bar{h}_i \) is estimated and subtracted from Eqn. (2) to obtain time series \( T_i' \):

\[ T_i' = (\Delta h_i(t_1), \ldots, \Delta h_i(t_T)), \]

with \( \Delta h_i(t_j) := h_i(t_j) - \bar{h}_i \). The purpose of the de-leveling is to avoid that height dominates consecutive clustering steps. An example of a de-leveled-input time series from PLS-KIJKDUIN is shown in Figure 1 in purple.

2.3 Cumulative Time series

Optionally, each time series is replaced by its cumulative version. That is, \( T_i \) in Eqn. 3 is replace by \( T_i^C \), defined as

\[ T_i^C = (\Delta_1, \Delta_1 + \Delta_2, \ldots, \Delta_1 + \Delta_2 + \ldots + \Delta_{T-1}, 0), \]

In Eqn.4, index \( i \) and height \( h \) is omitted for presentation purposes. Note that indeed

\[ \Delta_1 + \Delta_2 + \ldots + \Delta_{T-1} + \Delta_T = 0, \]

in Eqn. 4 as the \( \Delta_j \)'s represent the deviations of the mean height of the time series, compare Eqn. 3. An example of a cumulative time series from PLS-KIJKDUIN is shown in Figure 1 in orange. The process of first de-leveling a time series and next replacing it by its cumulative version, effectively brings down the length of the feature vector by one, as the last entry of Eqn. 5 is always equal to zero, and is therefore not contributing to the clustering.

2.4 Clustering

A large variety of un-supervised classification or clustering techniques is available, e.g. (Liao, 2005). In this work an easy to
implement and straightforward method is used, k-means clustering. One possible disadvantage of this method is that it requires the user to on forehand specify the number \( k \) of required clusters. K-means clustering assigns iteratively each of the \( T \)-dimensional time series to one of \( k \) cluster centers in \( \mathbb{R}^T \). After each iteration, cluster centers are recomputed based on the assigned time series, (Bishop, 2007). The method is initialized using random cluster centers, but is sensitive for the initiation. That is, different runs do not necessarily result in the same clusters.

The input of a run of k-means clustering is in our case a set of \( n \) time series of \( T = 6 \) epochs, ALS-TEXEL case, and \( T = 30 \) epochs, PLS-KIJKDUIN case. K-Means clustering is used to divide the \( T \)-dimensional feature space in \( k \) clusters. Note that the locations and the (mean) elevation of the different time series are not given to the \( k \)-means algorithm.

### 2.5 Possible extensions

Several, some silent, assumptions and choices were made in the above. The major ones are the following:

- Gridding or interpolation method
- Time series representation
- No-data handling
- Clustering metric
- Cluster method

It would be interesting to let a gridding method also incorporate the presence of vegetation by considering for example multiple vertical levels. Also the quality individual height estimates could be incorporated using a weighting scheme. Notably in the case of longer time series, it should be considered to fill missing data in certain epochs using some method of spatio-temporal interpolation. In this case the \( T \)-dimensional Euclidean metric was used within the k-means clustering. Except for considering possible alternative clustering methods, also the influence of the choice of metric should be assessed.

The PLS-KIJKDUIN data has been processed using Python, while ALS-TEXEL has been processed in Mathematica. Both data sets were processed on a standard desktop.

### 3. DATA DESCRIPTION

The proposed methodology is tested on two different data sets from two different acquisition systems. The first data set is acquired by a permanent, terrestrial, laser scanner, the second data set is acquired by airborne laser scanning. Both data sets sample typical Dutch sandy beach environments, characterized by a rather flat sandy beach, flanked by a first row of dunes covered by marron grass.

<table>
<thead>
<tr>
<th>Year</th>
<th>Nr. points</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>206 377</td>
</tr>
<tr>
<td>1999</td>
<td>129 151</td>
</tr>
<tr>
<td>2000</td>
<td>590 907</td>
</tr>
<tr>
<td>2001</td>
<td>566 852</td>
</tr>
<tr>
<td>2002</td>
<td>590 512</td>
</tr>
<tr>
<td>2003</td>
<td>454 447</td>
</tr>
</tbody>
</table>

Table 1. Number of input points per yearly epoch for the ALS-TEXEL data. Note that the number of points in the early years was considerably lower.

### 3.1 Permanent laser scan data

The first data set has been collected by a permanently installed Rieg-VZ 2000 scanner. The scanner was stably fixed on the top...
of a hotel roof in Kijkduin, The Netherlands, (Vos et al., 2017). The effective scan range of the Riegl-VZ 2000 in this setup is about 1 km and the field of view of the scanner contained some buildings, dunes, paths to the sandy beach and the beach itself. The first scan of the time series considered, is shown on the right in Figure 2. The scanner obtained one high-resolution scan every day and collected hourly low-resolution scans. The data consists of 31 point clouds in Cartesian xyz coordinates of about 2.5 million points each. In this work, one scan for each day of January 2017 is considered, approximately acquired at low tide. Only the data set of January 9, 2017 was discarded, due to bad weather. As a consequence, each time series consists of 30 epochs. Time differences between consecutive acquisitions are assumed to be exactly one day. The time gap of two days between the two consecutive epochs of January 8 and January 10 is not considered in any special way. After removing grid points with data missing in one or more epochs, a total of 23 508 complete time series entered the cluster analysis.

3.2 Repeated airborne laser scan data

The second data set consists of a set six yearly epochs of early airborne laser scan data, sampling the South-West of the Dutch island of Texel, compare also (Lindenbergh, Hanssen, 2003). The data set covers the period 1998-2003 and an area of about 1 km along shore and 0.5 km cross shore is considered in this work. The number of points available in each year is given in Table 1. In the early years, the number of points was relative low, which means that there will be considerable correlation between spatially near grid points after the epoch wise interpolation step. The time step between consecutive epochs is assumed to be exactly one year. A visualization of the last airborne scan, from 2003, is shown on the left in Figure 2. Back in 2003, changes were detected in this area using a classic geodetic approach of fitting low parameter kinematic models and comparing the quality of fit, while taking the number of parameters into account. After removing grid points having no elevation in at least one epoch, 221 459 complete time series entered the cluster analysis.

4. CLUSTERING RESULTS

The results are divided over three paragraphs. First, some initial results are shown for PLS-KIJKDUIN, obtained using a subset of the full time-series data set. Next, full results for PLS-KIJKDUIN are shown followed by the results for ALS-TEXEL.

4.1 Initial clustering results, PLS-KIJKDUIN

Some initial results are shown in Figure 3. For efficiency reasons, a subset of 10 000 spatial locations was randomly sampled from the one month time series of the Kijkduin permanent laser scan data. These time series, each consisting of 30 heights were de-leveled and given to the k-means implementation, using \( k = 10 \). Although the clustering method didn’t ‘know’ the actual heights of different locations, it gave back clusters that can be partially ordered by height. Indeed points on the beach (green zone) where automatically separated from points in the dunes (yellow zone). Still, the method also automatically distinguished time series from the dynamic intertidal area (dark-green) from the more stable area higher on the beach (light-green).

4.2 Full clustering results, PLS-KIJKDUIN

The full clustering results for PLS-KIJKDUIN are shown in Figure 4. Results are shown for k-means, applied to the de-leveled time series. Note that clustering results are not unique.
Figure 4. K-means clustering results for PLS-KIJKDUIN. The left column shows one result for k=5. On the top, time series locations are colored by cluster. On the bottom, for each cluster, the mean of all de-leveled time series belonging to that cluster is shown. The same applies for the right column, but here one result for k=10 is shown. The colors in the time series plots correspond to those in the spatial plots.

In our case k-means clustering is run using random initialization of initial cluster centers. If the implementation uses different random seeds for different runs, different outcomes are possible. In practice, we see that in all cases there are a few different outcomes that occur frequently.

A comparison to Figure 2 shows that resulting clusters for both k=5 and k=10 indeed are not dominated by height, but by the shape of the time series. This is most clear in the area on the right of the plots in Figure 4. For k=5 there are two dominant clusters, colored yellow and purple, on the dune slope connecting beach and boulevard. The mean de-leveled yellow time series, shown in the bottom left plot, appears more stable than the purple time series, but differences are subtle. Compared to k=5, the results for k=10 further differentiate between these subtle patterns of change. For example, the dark blue cluster on the lower part of the beach for k=5 is divided into two clusters colored in different shades of green for k=10.

The results also show that outlying time series affect the results. They either affect the centers of large clusters when lower values of k are chosen, or result in relatively small clusters of their own, when a big higher value of k is chosen. For example, in the k=10 case, there is a yellow cluster that appears in the time series plot, that is difficult to find back in the spatial plot, as it has too little points.
Figure 5. K-means clustering results for ALS-TEXEL. The top figure shows one result for $k=5$. The bottom figure shows one result for $k=10$. In both cases, time series locations are colored by cluster. In the middle, for each cluster, the mean of all time series belonging to that cluster is shown. The middle left shows the mean time series corresponding to the $k=5$ case of the top, while the middle right corresponds to the $k=10$ case at the bottom. The colors in the time series plots correspond to those in the spatial plots.


4.3 Clustering results, ALS-TEXEL

<table>
<thead>
<tr>
<th>k=5</th>
<th>k=10</th>
<th>k=5 (run 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>113 930</td>
<td>76 206</td>
<td>127 662</td>
</tr>
<tr>
<td>42 065</td>
<td>30 983</td>
<td>66 229</td>
</tr>
<tr>
<td>40 025</td>
<td>29 639</td>
<td>20 068</td>
</tr>
<tr>
<td>18 872</td>
<td>14 746</td>
<td>7 487</td>
</tr>
<tr>
<td>6 567</td>
<td>12 103</td>
<td>1 3</td>
</tr>
<tr>
<td>1 205</td>
<td>8537</td>
<td>1 3</td>
</tr>
<tr>
<td>1 195</td>
<td>7322</td>
<td>1 3</td>
</tr>
<tr>
<td>1 185</td>
<td>3859</td>
<td>1 3</td>
</tr>
</tbody>
</table>

Table 2. Number of time series per cluster for displayed cluster results for ALS-TEXEL, for k=5 (2x) and k=10.

The full clustering results for ALS-TEXEL are shown in Figure 5. In this case, k-means was applied to the de-leveled and cumulated time-series. Table 2 gives the number of time series per cluster, corresponding to the k=5 and k=10 runs in Figure 5. Note, again, that cluster results are not unique, but depend on the random seed initiation. The numbers 1 to 6 in the time series plots correspond to the six different epochs, one for each year from 1998 until 2003.

In contrast to the presentation of the results for PLS-KIJKDUIN, the mean time series in Figure 5 are the mean of the time series including elevations. For k=5, the resulting clusters appear to be sorted by elevation, but the k=10 case demonstrates that mean time series from different clusters do cross each other. A first conclusion is therefore that different morphological processes occur for points at different elevation. The results presented here do match previous results presented in (Lindebergh, Hansen, 2003) and (Vosselman, Maas, 2010) although there partly different years were analyzed. In all cases, elevation change occurs at the top of the first dune row, corresponding to the red, and less clear for the purple cluster for k=5, and for the purple and dark blue cluster for k=10. Previously, only a linear trend was considered, while the results shown here, indicate that some acceleration in accumulation took place between epochs 3 and 4, i.e. between 2000 and 2001. These more subtle change patterns get lost when estimating low parameter kinetic processes. Patterns match those as shown in (Vosselman, Maas, 2010), Figure 7.7, but the results here, once again show more subtle details.

5. DISCUSSION

In this chapter several issues are discussed.

5.1 Effect of elevation on clustering

Figure 6 shows what happens when elevation is not removed from the time series. In this case, the original PLS-KIJKDUIN time series were clustered using k = 5. The result is strongly dominated by elevation, compare also Figure 2.

![Figure 6](image-url)

Figure 6. Clusters resulting from including original elevations in time series are dominated by height.

5.2 Effect of random initiation

Figure 7 shows the result of a second run, using exactly the same input data and settings as used to produce Figure 5. In addition, in Table 2, the sizes of the resulting clusters are given for this second run. These sizes differ considerably. The fifth cluster for example consisted of 6 567 time series in the first run, and of only 13 time series in the second run. Still, the overall morphological impression of Figure 7 and Figure 5 is similar. When running k-means, typically only a few clustering configurations emerge, when using random initialization. The results in Figures 2 and 5 indicate that two different outcomes both may be interesting, as they reveal different significant patterns, dominated sometimes by outliers, sometimes by morphology. How such information could be exploited, will be studied in future research.

5.3 Computational efforts

The computational efforts required to run the proposed workflow are limited. Basically, the workflow consists of two steps: gridding and clustering. For gridding different approaches exist, but many require not more effort than sorting the n input that one should be careful when using k-means in combination with outliers: while the results for k=5 appear valid and beyond suspicion, the results for k=10 are clearly affected by outliers. This means however that these outliers are also affecting one or more clusters for k=5.

Finally note that the time series on the beach, i.e. the approximately first 200 m from the right in the k=5 and k=10 plots in Figure 5 show a largely consistent (between k=5 and k=10) decomposition into clusters. Previous analysis showed that in this areas varying change patterns are expected due to sand supple- tions. Patterns match those as shown in (Vosselman, Maas, 2010), Figure 7.7, but the results here, once again show more subtle details.
In this work un-supervised clustering of elevation time series is proposed and tested to distinguish different patterns of temporal change in multiple repeated laser scan data. The clustering is tested on one airborne and one permanent terrestrial laser scan data set. The airborne data set consists of six consecutive yearly epochs, while the terrestrial data sets consists of 30 daily epochs. For the clustering, k-means clustering is chosen, using Euclidean distances on de-leveled and in one case cumulative epochs. For the clustering, k-means clustering is chosen, using Euclidean distances on de-leveled and in one case cumulative epochs.

The experiments show that the proposed work-flow is easy to implement and powerful in revealing and distinguishing subtle patterns of change. There are also some issues. The method is in a way unstable, in the sense that running it using a random initialization may result in different outcomes, depending on the random seed that is used. In addition, the current work-flow requires full time series without any missing epochs. A known disadvantage of k-means clustering is that the number of clusters has to be specified by the users, and is not based on the spread of the data.

Therefore the next step is to systematically consider which clustering method is most suitable. In addition, the number of epochs considered in this first test was limited to 30, while novel permanent laser scan data sets consists of thousands of consecutive epochs, (Vos et al., 2017).

6. CONCLUSIONS

In this work un-supervised clustering of elevation time series is proposed and tested to distinguish different patterns of temporal change in multiple repeated laser scan data. The clustering is tested on one airborne and one permanent terrestrial laser scan data set. The airborne data set consists of six consecutive yearly epochs, while the terrestrial data sets consists of 30 daily epochs. For the clustering, k-means clustering is chosen, using Euclidean distances on de-leveled and in one case cumulative time series.

The experiments show that the proposed work-flow is easy to implement and powerful in revealing and distinguishing subtle patterns of change. There are also some issues. The method is in a way unstable, in the sense that running it using a random initialization may result in different outcomes, depending on the random seed that is used. In addition, the current work-flow requires full time series without any missing epochs. A known disadvantage of k-means clustering is that the number of clusters has to be specified by the users, and is not based on the spread of the data.

Therefore the next step is to systematically consider which clustering method is most suitable. In addition, the number of epochs considered in this first test was limited to 30, while novel permanent laser scan data sets consists of thousands of consecutive epochs, (Vos et al., 2017).

7. ACKNOWLEDGMENTS

Part of this work belongs to the CoastScan project, an Open Technology Programme with project number 16352, which is (partly) financed by the Dutch NWO Domain Applied and Engineering Sciences. The Dutch Public Works Dept. is thanked for making the ALS-TEXEL data available for this research. Finally, Katharina Anders is thanked for contributing considerably to the discussions on the processing of the PLS-KIJKDUIN data.

REFERENCES


