UC Berkeley
International Conference on GIScience Short Paper Proceedings

Title
A Principal Curve-based method for Geospatial Data Smoothing

Permalink
https://escholarship.org/uc/item/8ng293vd

Journal
International Conference on GIScience Short Paper Proceedings, 1(1)

Authors
Xiliang, Liu
Feng, Lu
Kang, Liu
et al.

Publication Date
2016

DOI
10.21433/B3118ng293vd

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

A Principal Curve-based method for Geospatial Data Smoothing

Xiliang Liu, Feng Lu, Kang Liu, Peiyuan Qiu, Li Yu, Mingxiao Li

State Key Lab of Resources and Environmental Information system,
Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences
{liuxl;luf;liukang;yul;limx}@lreis.ac.cn

Abstract

We propose a principal curve-based method for geospatial data smoothing. Firstly we test its performance with traditional approaches using floating car data (FCD). Secondly we evaluate its robustness with spatial-temporal dependence using Spearman rank correlation analysis. Final results show that the proposed method not only takes precedence over traditional methods (Mean and Median) in accuracy (about 10%-15% higher in RMSE), but also performs more robust, showing a distinct changing trend of the original data. These findings demonstrate the feasibility of the principal curve-based method in geospatial data smoothing.

Keywords: Principal curves, Data smoothing, Robustness

1. Introduction

Nowadays various GPS-equipped sensors, such as operating vehicles (taxicabs, probe cars, buses, private cars, etc.), mobile phones, wearable devices and so on, have become the mainstream in the research of GIScience and many other location based services (LBS) because of the cost-effectiveness and flexibility compared with other data sources. However, these geospatial data collected from GPS devices cannot be utilized directly owing to: (1) the sampling interval in most cities is low-frequency due to transmission bandwidth, energy consumption and storage pressures, and (2) the spatial-temporal distribution of these GPS-equipped devices among a city or a given region is heterogeneous. To further mine these data, the data sparseness, data missing and noise problem make geospatial data smoothing an unavoidable step.

Previous studies mainly focus on parametric approaches including Kalman filter, particle filter, piecewise linear (PWL) curves, and so on. These methods can effectively deal with data noise problem, but behave unsatisfactory with data sparseness and data missing problems. Traditional Mean and Median are also conducted by using current and near-past records from a historical perspective. However, the prerequisite of Gaussian distribution in most cases cannot be satisfied so that the traditional Mean and Median methods can only be employed in linear systems.

In this paper, we propose a principal curve-based method in geographical data interpolation. We evaluate its performance using floating car data (FCD), and analyze its robustness with Spearman's rank correlation analysis. A series of experiments demonstrate its feasibility for geospatial data smoothing.

2. Methodology

Principal curves give a summarization of the data in terms of a 1-*d* space nonlinearly embedded in the data space (Hastie and Stuetzle 1989). The original definition of a principal curve $f(t) = (f_1(t), ..., f_d(t))$ relies on the self-consistency property of principal components, following these requirements: (1) f does not intersect itself; (2) f has finite length inside any bounded subset; (3) f is self-consistent.

We design the iterative strategy for principal curves as follows:

```
**Algorithm 1**. PrincipalCurveIteration
**Input**: Time-labelled geographical data list $x$;
        Initial value formula $f$;
**Output**: principal curve list $t_f(x)$
1: Initialization.
   The initial principal curve $f^{(j)}(t)$ is defined as the first line principal component, here $j=1$;
2: Projection.
   $\forall x \in R^d$, calculate the $t_f(x)$ in Equation 2 using Euclidean distance.
3: Expectation.
   Based on the self-consistent character, the first principal curve is re-calculated as follows:
      $f^{(j)}(t) = E[X \mid t_{f^{(j)}}(X) = t]$      (3)
4. Adjustion.
   If $1 - \frac{\Delta(f^{(j+1)})}{\Delta(f^{(j)})} < \varepsilon$, the iteration is stopped;
   Else $j = j+1$, and go to step 2
```

To evaluate the performance of this proposed method, we employ two means: the smoothing accuracy and the spatial-temporal dependence using root mean square error (RMSE) and Spearman's rank correlation analysis.

3. Experiments

3.1 Smoothing accuracy analysis

We employ intersection delay data (Liu *et al*, 2013) which are derived from Beijing's FCD to analyze the smoothing accuracy. 400 main intersections from all the 14,614 ones are selected to represent the skeleton of Beijing's road network. We record the road ID, the driving directions, the time interval ID and the intersection delays. The whole process is as follows:

(1) For a given intersection, the total turn delay records is expressed as $Td = \{tde_1, td_2, \ldots, td_n\}$.Here n stands for the number of driving directions. For a given driving direction i ($i=1,2,..,n$), the intersection delay records are $td_i = \{td_{i1}, td_{i2}, \ldots, td_{im}\}$, where m stands for the record number in a given time slot. The time slot starts from 1 to 96, the span of time interval is 15 minutes in the original intersection delay dataset;

(2) Test if the intersection delays follow normal distribution in this time span. If so, go to (4), or else go to(3);

(3) Smooth the given intersection delay records based on the proposed principal curve method (Algorithm 1).

(4) Average all the intersection delays as the final turn delay value for this given time slot.

We also employ two classical approaches, Mean and Median. In order to demonstrate the performance for the real dynamic situations, we apply these three methods for the intersection delay smoothing for a given driving direction of a selected intersection in Figure 1.

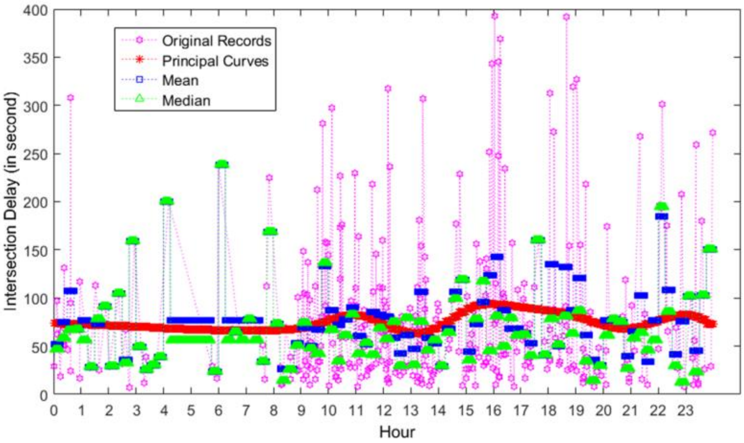

Figure 1. Results comparison (in a whole day)

During all the experiments among these 400 main intersections, the proposed principal curve-based method generally surpasses the traditional Mean and Median methods by 10%-15% in root mean square error (RMSE) for a given FCD file, more higher during the peak hours and for the marginal intersections which have fewer taxicab records.

3.2 Spatial dependence analysis

We utilize road travel time dataset of Beijing's road network ranging from February to June in 2012. In total, the average length of the road segments is 309.7 meters. The sampling interval for this dataset is 5 minutes, with 288 time slots in a day. The main problems of this dataset exist in that the records of some road segments are not complete due to heterogeneous distribution of GPS devices. Furthermore, some records are obviously abnormal during a given time interval.

We put the principal curve-based method, the mean and median into the smoothing of the road travel time dataset with the same parameters in Algorithm 1. In the implementation of Mean and Median, the time window size is set as 5 according to Liu *et al*. (2013) so as to keep the details of the original data and satisfy the smoothing requirements. We perform the Spearman's rank correlation analysis for these three methods and fit the original data with linear regression. For simplicity, we compare the fitted values between three different methods, as shown in Figure 2. In Figure 2, the fitted values of Spearman's rho between each results and the original data all elevate when the road' length increases, and the correlation between the results of principal curves and the original data behaves the strongest compared with the other two traditional ones.

3.3 Temporal dependence analysis

We design the experiment for temporal dependence analysis based on the Shanghai's high-frequency FCD. The average sampling interval is 5 seconds, ranging from 2012.10.1 to 2012.12.31. A typical intersection is selected with 1,509,906 trajectories and 1,006,708,483 GPS observations. For each driving direction, we calculate the intersection delays under different sampling intervals (5s, 10s, 15s, 30s, 60s, 90s, 120s). In order to quantify the temporal dependence among different sampling intervals, we first take the all intersection delay records of a given driving direction as a whole time series list. Then we employ different methods, including principal curves, Mean and Median. After the smoothing, we estimate the Spearman's rank correlation coefficient ρ between different results and the original data of a specific driving directions under different sampling intervals (i.e. 5s, 10s, 15s, 30s, 60s, 90s, and 120s). Finally, we average all the 12 driving directions' Spearman's ρ

under different sampling intervals. Figure 3 illustrates the results of temporal dependence analysis.

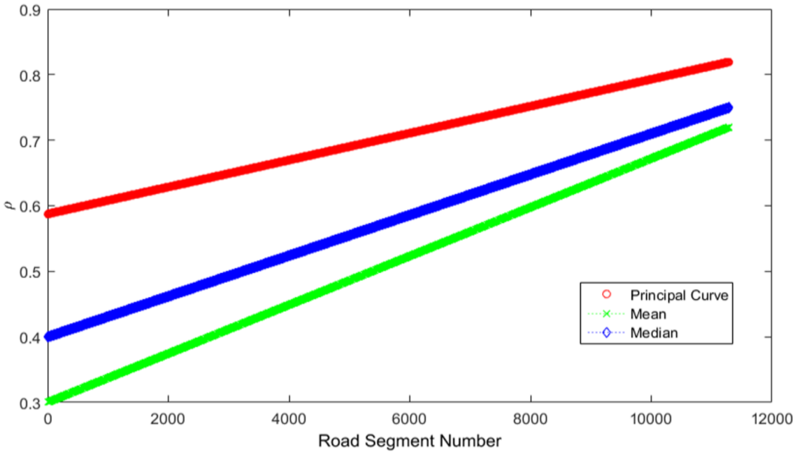

Figure 2. Spatial dependence comparison between fitted values of Spearman's rho

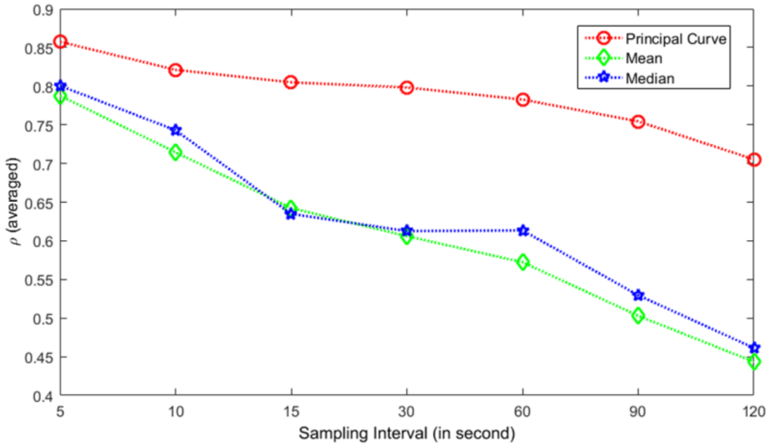

Figure 2. Temporal dependence comparison between fitted values of Spearman's rho

4. Conclusions

We propose a principal curve-based method for geospatial data smoothing. The proposed method not only takes precedence over traditional methods (Mean and Median) in accuracy (about 10%-15% higher in RMSE), but also performs more robust in dealing with data sparseness, data missing and noise problems, showing a promising feasibility in geospatial data smoothing.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant No. 41271408), the National Hi-tech Research and Development Program of China (Grant No. 2013AA120305) and China Postdoctoral Science Foundation funded project (Grant No. 2015M581158).

References

Hastie T, Stuetzle W, 1989, Principal curves. *Journal of the American Statistical Association*, 84(406):502-516.

Liu X L, Lu F, Zhang H C, Qiu P Y, 2013, Intersection delay estimation from floating car data via principal curves: a case study on Beijing's road network. *Frontiers of earth science*, 7(2):206-216.