Accident Analysis and Prevention 71 (2014) 29–37

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Denoising traffic collision data using Ensemble Empirical Mode Decomposition (EEMD) and its application for constructing Continuous Risk Profile (CRP)

Nam-Seog Kim[a,1], Koohong Chung[b,c,*], Seongchae Ahn[d,2], Jeong Whon Yu[d,3], Keechoo Choi[d,2]

[a] Department of Electrical Engineering and Computer Science, University of California at Berkeley, United States
[b] School of Civil, Environmental and Architectural Engineering, Korea University, Anam-Dong, Seongbuk-Gu, Seoul 136-713, Republic of Korea
[c] California Department of Transportation Operations, Highway Operations, 111 Grand Ave, Oakland, CA 94623, United States
[d] Department of Transportation Engineering, Ajou University, San 5, Woncheon-dong, Yeongtong-gu, Suwon 443-749, Republic of Korea

ARTICLE INFO

Article history:
Received 17 December 2013
Received in revised form 6 May 2014
Accepted 6 May 2014

Keywords:
Empirical Mode Decomposition
Intrinsic Mode Function
Continuous Risk Profile
Traffic collision

ABSTRACT

Filtering out the noise in traffic collision data is essential in reducing false positive rates (i.e., requiring safety investigation of sites where it is not needed) and can assist government agencies in better allocating limited resources. Previous studies have demonstrated that denoising traffic collision data is possible when there exists a true known high collision concentration location (HCCL) list to calibrate the parameters of a denoising method. However, such a list is often not readily available in practice. To this end, the present study introduces an innovative approach for denoising traffic collision data using the Ensemble Empirical Mode Decomposition (EEMD) method which is widely used for analyzing nonlinear and nonstationary data. The present study describes how to transform the traffic collision data before the data can be decomposed using the EEMD method to obtain set of Intrinsic Mode Functions (IMFs) and residue. The attributes of the IMFs were then carefully examined to denoise the data and to construct Continuous Risk Profiles (CRPs). The findings from comparing the resulting CRP profiles with CRPs in which the noise was filtered out with two different empirically calibrated weighted moving window lengths are also documented, and the results and recommendations for future research are discussed.

Published by Elsevier Ltd.

1. Introduction

Government transportation agencies have long monitored the locations of traffic collisions that occur on their roadways in an effort to identify high collision concentration locations (HCCL) by ranking roadway sites based on different types of statistics observed within a site's boundaries. Such HCCL detection efforts involve grouping roadways based on their features (i.e., grade and number of lanes) (Federal Highway Administration, 2010), developing Safety Performance Function (SPF) for each roadway group (Kwon et al., 2013; Tegge et al., 2010), and predefining boundary locations of potential HCCL sites before applying network screening procedures such as the Sliding Moving Window (SMW) and Peak Searching (PS) methods (Federal Highway Administration, 2010; Kwon et al., 2013).

The existing approaches have been instrumental in prioritizing sites for safety investigation and resource allocation for implementing safety countermeasures. However, existing network screening procedures, such as SMW and PS, suffer from a high false positive rate (i.e., requiring safety investigation of sites where it is not needed) since these methods do not consider the spatiotemporal relationship among the traffic collision data, and traffic collision data used in the analysis is grouped following a heuristic approach (Kwon et al., 2013; California Department of Transportation, 2002). Predetermining boundary locations based solely on the site's physical characteristics implicitly assumes that traffic collision causative factors reside within the predefined boundary. Failing to address the spatiotemporal relationship among the traffic collisions and the

* Corresponding author at: School of Civil, Environmental and Architectural Engineering, Korea University, Anam-Dong, Seongbuk-Gu, Seoul 136-713, Republic of Korea. Tel.: +1 510 622 5429; fax: +1 510 286 4561.
E-mail addresses: namseog@eecs.berkeley.edu (N.-S. Kim), koohong_chung@dot.ca.gov (K. Chung), citadel@ajou.ac.kr (S. Ahn), jeongwhon@ajou.ac.kr (J.W. Yu), keechoo@ajou.ac.kr (K. Choi).
[1] Tel: +1 510 543 4929.
[2] Tel: +82 10 8653 5202.
[3] Tel: +82 31 219 1650.

http://dx.doi.org/10.1016/j.aap.2014.05.007
0001-4575/Published by Elsevier Ltd.

implicit assumption of collision causative factor locations can contribute to an increase in the high rate of false positives (California Department of Transportation, 2002; Chung et al., 2009). To address this issue, Chung et al. (2009, 2011) proposed the Continuous Risk Profile (CRP) as an alternative to SMW and PS.

The CRP method (Kwon et al., 2013; Chung et al., 2009, 2011; Jang et al., 2009) first filters out random fluctuations in data using a weighted moving average technique (Ljung, 1999). Next, it allows traffic collision data and SPFs to determine the endpoints of a site. Finally, the data within the site's boundaries are used to estimate the statistics used for ranking sites for safety investigation (California Department of Transportation, 2002; Washington et al., 2014). Empirical evaluation of the CRP approach confirmed that it improves performance of detecting high collision concentration locations over the existing SMW and PS methods (Kwon et al., 2013)—CRP markedly reduced the false positive rate while its false negative rate was comparable to SMW and PS.

Application of the CRP method requires that one of its parameters, the moving window size used to filter out random noise in the data, be empirically calibrated. Using too large a window size can result in altering the locations of the peaks (i.e., high collision concentration locations), while using too small a window size may compromise the performance of the high collision detection procedure. Chung et al. (2011) demonstrated how the proper range of window length can be determined, however such calibration may not be feasible where a true HCCL list is not readily available.

To this end, the present study developed a systematic method of denoising traffic collision data to calculate CRPs without the need for a true HCCL list by employing the Ensemble Empirical Mode Decomposition (EEMD) method (Wu and Huang, 2009). The proposed approach decomposes the transformed traffic collision data into a set of Intrinsic Mode Functions (IMFs) and residue. Next, it separates noise from the relevant signal based on the statistical significance of each IMF (Huang et al., 1998; Wu and Huang, 2004) to construct a CRP. The descriptions of empirical mode decomposition (EMD) and EEMD, the denoising process for constructing CRP and reports on the findings are explained in Section 2. The paper ends with brief concluding remarks in Section 3.

2. Background

EMD is a data adoptive method for analyzing nonlinear and non-stationary data and has been widely adopted in number of different fields (Yu et al., 2005; Ditommaso et al., 2012; Hariharan et al., 2006) because of its potential advantages over similar existing methods such as the Short Time Fourier Transform (STFT) and wavelet approaches (Huang et al., 1998). However, EMD may not function properly if the data does not meet certain conditions (Huang et al., 1998; Flandrin et al., 2005), as in the case of traffic collision data. Therefore, the original traffic collision frequency data were transformed prior to applying the EMD method to decompose the data into set of IMFs and residue through a process called sifting (Huang et al., 1998). Noise is introduced during the decomposition process to prevent mode mixing (Wu and Huang, 2009) from contaminating the information embedded in IMFs. This noise-assisted EMD method is called EEMD.

2.1. Transforming traffic collision data prior to applying the EMD method

The EMD method was first developed by Huang et al. (1998) to analyze nonlinear and non-stationary data. The objective of the method is to identify the intrinsic oscillatory modes by their characteristic frequencies in the data empirically, and to then decompose the data accordingly into set of IMFs. Since the method is data adaptive, it does not require identification of a stationary period, or selection of a wavelet with which to analyze the data. The decomposition method, however, cannot be directly applied to the traffic collision frequency data as explained below and illustrated in Fig. 1.

Fig. 1(a) shows traffic collision frequency, $t(d)$, observed with respect to distance, d, along I-880S located near the San Francisco

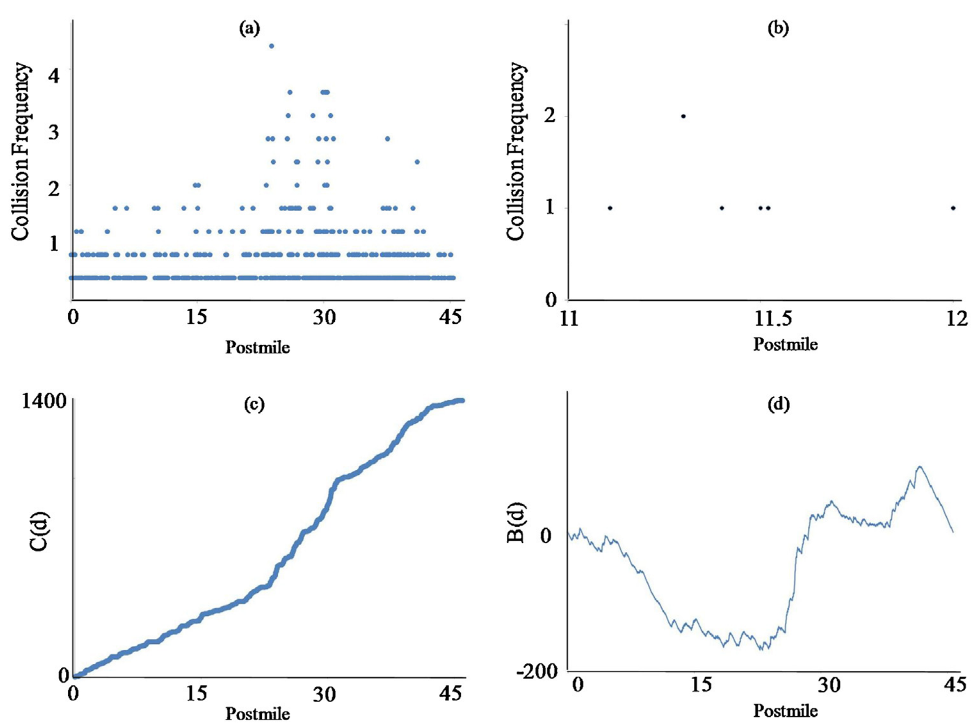

Fig. 1. Transformation of traffic collision data. (a) Traffic collision frequency reported along I-880S in 2008 between postmiles 10 and 20. (b) Traffic collision frequency reported along I-880S in 2008 between postmiles 11 and 12. (c) Cumulative traffic collision frequency, $C(d)$, (d) transformed $C(d)$