Transportation Research Part C 69 (2016) 375–385

Contents lists available at ScienceDirect

Transportation Research Part C

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

A mathematical programming technique for matching time-stamped records in logistics and transportation systems

L. Douglas Smith [a,*], Jan Fabian Ehmke [b]

[a] *University of Missouri-St. Louis, One University Blvd., St. Louis, MO 63121, USA*
[b] *Freie Universität Berlin, Garystr. 21, D-14195 Berlin, Germany*

ARTICLE INFO

Article history:
Received 18 January 2016
Received in revised form 31 May 2016
Accepted 7 June 2016
Available online 29 June 2016

Keywords:
Data mining
Preprocessing
Mathematical programming
Time-stamped records
Automated matching

ABSTRACT

Time-stamped data for transportation and logistics are essential for estimating times on transportation legs and times between successive stages in logistic processes. Often these data are subject to recording errors and omissions. Matches must then be inferred from the time stamps alone because identifying keys are unavailable, suppressed to preserve confidentiality, or ambiguous because of missing observations. We present an integer programming (IP) model developed for matching successive events in such situations and illustrate its application in three problem settings involving (a) airline operations at an airport, (b) taxi service between an airport and a train station, and (c) taxi services from an airport. With data from the third setting (where a matching key was available), we illustrate the robustness of estimates for median and mean times between events under different random rates for "failure to record", different screening criteria for outliers, and different target times used in the IP objective. The IP model proves to be a tractable and informative tool for data matching and data cleaning, with a wide range of potential applications.

1. Introduction

Analysis and modeling of logistics and transportation systems frequently involve the integration of temporal data from multiple sources for sequential steps in a process. From the blended data, time intervals between successive events are determined. Such measures often represent important aspects of system performance and provide essential information for constructing reliable models of system behavior. They are used to produce parameters for probability distributions used in simulation models, parameters for optimizing models, and metrics for performance dashboards. Decisions based on the resulting statistics (or on models derived from them) can be critical determinants of future system performance.

In many situations, there are lapses or errors in the recording of information – either from human error or mechanical failure. Further, the information systems used to collect the data may or may not have a common key for identifying the individual entities. Gordon et al. (2008) tested the reliability of time stamps from both active recorders (individuals who perform some action such as entering a code on some device) and passive recorders (where events are recorded incidentally by readers such as wireless sensors) in a hospital emergency department. They found that both were fraught with systematic errors that prevented accurate measures of time intervals between successive activities. Wang and Liu (2005) describe an information architecture for managing temporal data collected by radio-frequency identification (RFID), acknowledge problems of duplicate readings and missing readings, mention the issue of blending transactional and locational data, but ignore

[*] Corresponding author.
E-mail addresses: ldsmith@umsl.edu (L.D. Smith), JanFabian.Ehmke@fu-berlin.de (J.F. Ehmke).

http://dx.doi.org/10.1016/j.trc.2016.06.007
0968-090X/© 2016 Elsevier Ltd. All rights reserved.

issues of data cleaning. Jeffery et al. (2006) report that reading failures as high as 30% occur for RFID equipment and discuss mathematical smoothing filters with adaptive time windows for interpolating values of missing observations. Gellrich et al. (2011), in constructing models of an internal logistics system, relied on logs of discrete events to determine transit times between points in the system. They found that "… the available data may be incomplete (e.g., information about the exact position of a lot in any point of time during the transport process is not recorded or details on transport control logic are missing), thus the system is only *partly observable*." They attributed the problem to recording errors and to difficulties in merging data from different sources with different formats. In this situation, they had missing information about intermediate movements of entities and had the benefit of a common identifier for the lots being transported. Their concern was with the estimation of elemental times in segments of alternative routes when the actual route differed from the normal path.

In this paper, we present an integer programming (IP) procedure for matching successive events in such situations and illustrate its application in three problem settings involving (a) airline operations at an airport, (b) taxi service between an airport and a train station, and (c) taxi services from an airport. The procedure allows an analyst to stipulate upper and lower limits on the inferred time intervals between the matched events (to eliminate outliers due to extraordinary events or recording errors) and to set a "target" for the normal time expected. By selecting different "target" values ranging from the lowest possible value to the highest possible value, the analyst may produce distributions for the inferred time intervals under assumptions ranging from pessimistic to optimistic interpretations of the data. The inferred distributions will depend, of course, on the screening criteria and frequency of recording lapses. With data from the airport taxi dispatching service (where a matching key was available), we illustrate the robustness of estimates for median and mean times between events under different random rates for "failure to record", different screening criteria for outliers, and different target times used in the IP objective. We thus address the following research questions:

1. Is the IP procedure sufficiently efficient to employ it as a data integration and data-cleaning tool?
2. Are estimates of mean or median values for inferred time intervals highly sensitive to the screening criteria and "target" values used in the matching process?
3. Can employment of the IP model with "target" values ranging between extremes offer a mitigating strategy for random loss of data?

2. Analytics for data integration and data cleaning

Before presenting our specific case studies and solution methodology, we briefly describe (and illustrate in Fig. 1) the general context within which our methodology is relevant. Technological advancements such as automatic sensors and ubiquitous connectivity have enabled the collection and storage of enormous amounts of data from logistics and transportation processes at low cost. The challenge is to integrate the data and transform them into information that can improve future logistics and transportation operations (Ehmke et al., 2009). Zhang et al. (2013) discuss issues and processes for filtering vehicle probe data and interpolating values when readings are sparse. Like Zhong et al. (2013), who proposed statistical models for replacing missing readings in traffic data, they deal with aggregate counts and speeds on a roadway in small time intervals, and not with the matching of successive stages in a vehicle's itinerary.

Data cleaning and data integration comprise the preprocessing of operational data from several sources for subsequent data mining. *Data cleaning* is required to make the data fit to their intended use (Berthold et al., 2010). It involves the identification of instances with missing values, the smoothing of noisy data, the identification and removal of outliers, and the detection of inconsistencies. Tan et al. (2009) and Berthold et al. (2010) provide an overview of methods for data cleaning, which are usually based on descriptive statistical analysis. *Data integration* refers to the conceptual and physical integration of data from several sources. Processes may be employed to ensure syntactic and semantic consistency and values may be normalized to cope with differences in metrics used (Han and Kamber, 2006). Laxman and Sastry (2006) describe "temporal

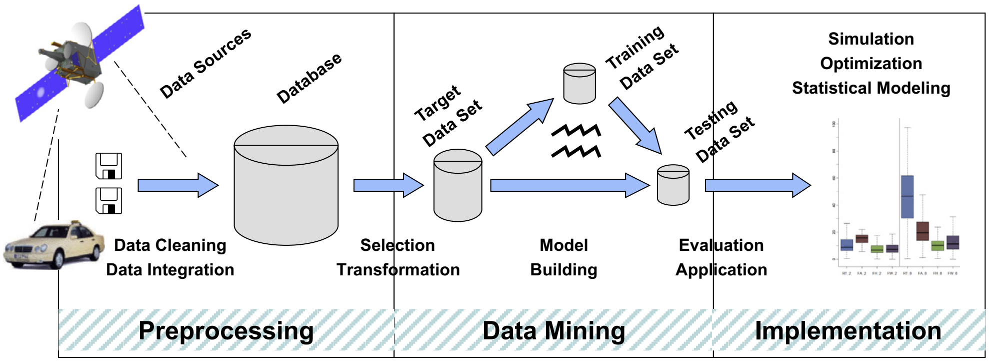

Fig. 1. Data cleaning and data integration in the context of the knowledge discovery process (adapted from Han and Kamber (2006)).