European Journal of Operational Research 000 (2018) 1–20

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Incomplete information imputation in limited data environments with application to disaster response

Kezban Yagci Sokat[a,*], Irina S. Dolinskaya[b], Karen Smilowitz[a], Ryan Bank[c]

[a] Department of Industrial Engineering and Management Sciences, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208, United States
[b] National Science Foundation, 2415 Eisenhower Avenue, Alexandria, VA 22314, United States
[c] National Insurance Crime Bureau,1221 W Cornelia Ave, Chicago 60657, United States

A R T I C L E I N F O

Article history:
Received 4 November 2016
Accepted 6 February 2018
Available online xxx

Keywords:
Humanitarian logistics
Incomplete information
Imputation
Limited data
2010 Haiti earthquake

A B S T R A C T

Following a major disaster, a field operations manager needs to deploy relief activities within the affected region. State-of-the-art humanitarian logistics models have been developed over the past decades to assist relief operations. However, while many models assume availability of information on infrastructure status, this is typically not the case in practice. Often, only partial information about infrastructure status is known. Utilizing the similarities in the known data via attributes, we develop a framework to impute incomplete information in limited data environments. We present an application of this framework to a past disaster, the 2010 Haiti earthquake. We build an ArcGIS model to automate the data collection and processing efforts to the extent possible. The study explores the impact of missing data, dispersion of missing data and imputation techniques used in approximating the incomplete information. Our results suggest that lower granularity yields better estimates of the unknown information above a threshold. We also develop publicly available test cases for the broader community.

1. Introduction

With the increasing number of large scale disasters, there has been more attention in the operations research community to develop humanitarian logistics models over the past decades. Often, this work presents deterministic models with the assumption of known data (Jiang, Yuan, Huang, & Zhao, 2012). However, complete data is generally not available or hard to gather and integrate in a disaster setting (Argollo, Bandeira, & Campos, 2013). It is even harder to obtain relevant data for transportation purposes despite the fact that transportation plays a critical role in humanitarian aid (Van Wassenhove, 2006). In a recent study, (Yagci Sokat, Zhou, Dolinskaya, Smilowitz, & Chan, 2016) discuss the availability of the infrastructural damage and quantify the available road damage information as 8% in the case of Typhoon Haiyan. There is uncertainty about the status of the remaining road segments, which implies that the majority of the information is incomplete.

The first update map provided by United Nations Institute for Training and Research's Operational Satellite Applications Programme (UNITAR/UNOSAT) three days after the 2010 Haiti Earthquake shows road status information about some road segments, highlighted as partially blocked or blocked (UNCS, 2010). However, the UNITAR/UNOSAT map does not provide information about which road segments are open and it is not clear whether the remaining unmarked roads are open or not. At the same time, field operations managers have to make urgent deployment decisions given this limited availability of information. This paper explores how we, as operation researchers, can assist field managers in making quick and efficient decisions in such settings.

To utilize every available piece of information to the greatest extent possible, we introduce an innovative approach that uncovers similarities between road segments with known road status and utilizes these similarities to fill in the knowledge gaps. We explore whether certain characteristics, defined as *attributes*, such as road type or distance to epicenter, can be used to understand the damage impact across the network. We conjecture that roads with similar attributes have a high correlation in their damage level. For example, if two road segments have a similar number of damaged buildings around each of them, and one of the road segments is known to be blocked, it is likely that the other road segment is also blocked. As part of this study, we develop a list of potential attributes that can be used to estimate road status information.

Utilizing new data sources (such as Open Street Map (2016) (OSM)) along with more traditional data sources for

* Corresponding author at: Department of Industrial Engineering and Management Sciences, Northwestern University, 2145 Sheridan Road C210, Evanston, IL 60208, United States.
E-mail address: kezban.yagcisokat@u.northwestern.edu (K. Yagci Sokat).

https://doi.org/10.1016/j.ejor.2018.02.016
0377-2217/

Please cite this article as: K. Yagci Sokat et al., Incomplete information imputation in limited data environments with application to disaster response, European Journal of Operational Research (2018), https://doi.org/10.1016/j.ejor.2018.02.016

road damage status and relevant attributes (such as portable data format (PDF) maps and shapefiles (SHP)- a geospatial vector data format for geographical information system), we develop a unifying framework to estimate incomplete information in limited data environments. Since much of the current publicly available transportation network data is disseminated in PDF formats, to make this information useful to decision models, the road status information has to be migrated to a digital geographical information platform such as ArcGIS ESRI (2016). This migration process requires extensive georectification (i.e., matching the PDF map to pre-existing road network), identifying the damage and manually entering this information into ArcGIS. ArcGIS is also critical for collecting the attribute information for individual road segments and preparing them for imputation algorithms. We build an ArcGIS model to automate these data gathering and processing steps to the extent possible. This model eliminates a significant amount of the manual work and enables faster data processing for use in future disasters.

Once the available data are captured and processed, we propose various imputation techniques to estimate the missing data. We utilize the properties of the available data structure and the unique characteristics of the post disaster environment to develop new algorithms: clustering combined with modified mean-and-mode and clustering combined with adjacent arcs. In addition to these unsupervised clustering methods, we also test supervised learning-based methods such as classification tree. We also employ global constant methods (optimistic, pessimistic, neutral and popularistic) for benchmarking, as they represent reasonable and easy to implement strategies.

We validate our framework with a case study based on the 2010 Haiti Earthquake. We estimate the status of roads after the disaster and demonstrate results with a high level of accurately predicted road damage. We provide key insights from this case study and discuss the applicability of this framework to other disasters. Our test cases are made publicly available for the broader community's use.

The remainder of this paper is organized as follows. Section 2 provides an overview of relevant literature focusing on the use of infrastructural data in humanitarian logistics models. Section 3 presents our framework for imputing incomplete information, while Section 4 presents the results from application of this framework to a past disaster case study. Section 5 discusses the implications of this framework for humanitarian response and broader areas. Finally, Section 6 summarizes the study with concluding remarks.

2. Related literature

2.1. Transportation network information in humanitarian logistics studies

Existing studies recognize the lack of infrastructure information following a disaster and the importance of this information to humanitarian logistics decisions (Chen & Miller-Hooks, 2012). However, to our knowledge, none of the prior work addresses incomplete information systematically. Below, we summarize the type of transportation network information discussed in the humanitarian logistics literature and examples of different types of data used. We should note that some studies assume Euclidean distances to overcome lack of data; however, we do not include such papers in our review. We also do not discuss the decision problem modeled, but rather focus on the modeling of the transportation networks.

To tackle the incomplete information in transportation networks, different elements of uncertainty are included in the humanitarian logistics models, such as travel time, availability, link capacity, and reliability (Liberatore, Pizarro, Blas, Ortuño, & Vitoriano, 2013). One approach to address uncertain travel times is the use of distributions. Shen, Dessouky, and Ordoñez (2009) use lognormal distribution to generate realizations of travel times. Huang, Fan, and Cheu (2007) model the travel times with a uniform distribution with a lower bound based on the free-flow speed and an upper bound based on a percent of noise added to the lower bound. Another method is to alter travel time by multiplying the travel times with a specific coefficient per scenario. Mete and Zabinsky (2010) create six scenarios based on the location of the earthquake, an expected earthquake in Seattle, USA triggered by Seattle fault or Cascadia fault, and the time of the day: working hours, rush hours and nonworking hours.

Link availability is also a popular approach to account for the uncertainty in the transportation network (Ahmadi, Seifi, & Tootooni, 2015; Akbari & Salman, 2016; Celik, Ergun, & Keskinocak, 2015; Clark & Culkin, 2007; Günneç & Salman, 2011). Link availability is generally modeled in two ways. The first option is to randomly destroy a certain percent of the roads (Ahmadi et al., 2015; Celik et al., 2015). For example, (Celik et al., 2015) randomly determine the blocked roads and vary the percent of blocked roads in the grid and ring networks. Another method is to assign probability of survival or failure for each link based on different scenarios or instances with respect to intensity and location of the disaster (Akbari & Salman, 2016; Günneç & Salman, 2011). (Günneç & Salman, 2011) adjust the link survival probabilities (the complement of link failure probability) with respect to disaster scenarios developed by experts for a possible earthquake in Istanbul, Turkey based on the magnitude and location of the earthquake (IMM, 2002).

Link capacity and reliability are also introduced to capture uncertainty in the transportation network (Barbarosoglu & Arda, 2004; Nolz, Semet, & Doerner, 2011; Rawls & Turnquist, 2012; Vitoriano, Ortuño, Tirado, & Montero, 2011; Yazici & Ozbay, 2007). Barbarosoglu and Arda (2004) employ random arc capacities that are generated based on the building damage scenarios developed by Erdik and Aydinoglu (2002) using the 1999 Marmara earthquake in Istanbul, Turkey. Rawls and Turnquist (2012) also use scenario dependent arc capacities to account for the damage in the transportation links for a possible hurricane in North Carolina, the United States (U.S.) with the hurricane scenarios coming from HAZUS (geographic information system-based natural hazard developed by the Federal Emergency Management Agency (FEMA, 2012)) and their previous research. Since the authors focus on models rather than the data, there is no discussion about how the arc capacities are obtained and scenarios are generated. Yazici and Ozbay (2007) develop a procedure to assign capacity loss probability for the specifically chosen links with application to Cape May network in New Jersey, U.S. The procedure runs through all arcs to identify one out of every three arcs and assign capacity loss probability to these arcs. (Yazici & Ozbay, 2007) follow a cell based approach, which captures the impact of geography in the capacity loss by assigning higher loss probabilities to cells that are at higher risk. For example, under a flooding case, authors allocate higher capacity loss probabilities to cells near a shore. Nolz et al. (2011) use the elementary catastrophe hazard of each arc calculated by catastrophe models for specific regions as risk of an arc to capture link reliability. In addition to reliability, (Vitoriano et al., 2011) model link security with the probability of a vehicle to be ransacked when traveling through an arc following the Haiti earthquake in 2010.

In summary, researchers use four common elements to model infrastructural damage information: traversing time, link availability, link capacity, and link reliability. However, most of these methods do not incorporate the available information from the field. In addition, there is no systematic scheme to account for