Computers, Environment and Urban Systems xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

# Big enterprise registration data imputation: Supporting spatiotemporal analysis of industries in China

Fa Li[a,c], Zhipeng Gui[b,a,c,*], Huayi Wu[a,c], Jianya Gong[b,a,c], Yuan Wang[a,c], Siyu Tian[b], Jiawen Zhang[b]

[a] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China
[b] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China
[c] Collaborative Innovation Center of Geospatial Technology, Wuhan, China

ARTICLE INFO

*Keywords:*
Geocoding
Missing values imputation
High Performance Computing
Industrial spatial distribution
Urban spatial structure
Short text classification

ABSTRACT

Big, fine-grained enterprise registration data that includes time and location information enables us to quantitatively analyze, visualize, and understand the patterns of industries at multiple scales across time and space. However, data quality issues like incompleteness and ambiguity, hinder such analysis and application. These issues become more challenging when the volume of data is immense and constantly growing. High Performance Computing (HPC) frameworks can tackle big data computational issues, but few studies have systematically investigated imputation methods for enterprise registration data in this type of computing environment. In this paper, we propose a big data imputation workflow based on Apache Spark as well as a bare-metal computing cluster, to impute enterprise registration data. We integrated external data sources, employed Natural Language Processing (NLP), and compared several machine-learning methods to address incompleteness and ambiguity problems found in enterprise registration data. Experimental results illustrate the feasibility, efficiency, and scalability of the proposed HPC-based imputation framework, which also provides a reference for other big georeferenced text data processing. Using these imputation results, we visualize and briefly discuss the spatiotemporal distribution of industries in China, demonstrating the potential applications of such data when quality issues are resolved.

## 1. Introduction

Big data with fine-grained street-level location and coordinates, as well as operating period and industrial category information can deepen and extend analysis of industrial spatial distributions, thereby promoting a deeper understanding of urban processes. The spatial distribution of various economic activities lies at the very heart of theories of urban spatial structure and is essential for rational urban and regional economic planning and policymaking (Li, Zhang, Chen, & Yu, 2015; Parr, 2014). However, due to the lack of complete, fine-grained micro-level enterprise or firm data, few studies have fully analyzed the spatial distribution of industries in China at multiple scales, from a temporally sensitive perspective, incorporating all kinds of enterprises (Watkins, 2014; Zhu & Chen, 2007). The local bureaus of Administration for Industry and Commerce (AIC) of China, are responsible for enterprise registration, supervision and administration, and protection of consumers' rights and interests (AIC, 2016). These regional bureaus record detailed operating information for each enterprise. Big enterprise registration data, collected from multiple regional bureaus of AIC of China, can enable and support spatial-temporal analysis of industries, if the data quality issues are resolved.

Incompleteness and address ambiguity are prominent quality problems of Chinese enterprise registration data. A typical registration record contains information of an individual enterprise, including enterprise name, address, registration date, industrial category, business scope, postcode, legal representative, and registered capital. Usually, these records are manually recorded and inputted into the system at local AIC offices. In this process, critical information is either overlooked or neglected, and therefore frequently missing from the database. For example, in our study, 43.64% of the data has no industrial category values. This information however, is imperative when executing a spatial distribution analysis of industrial categories and industries. Approximately 30% of the records only have a street-level address but do not include the province or city to which it belongs. This address ambiguity problem is defined as the missing Administrative Division (AD) information problem, seriously impeding effective

[*] Corresponding author.
*E-mail address:* zhipeng.gui@whu.edu.cn (Z. Gui).

https://doi.org/10.1016/j.compenvurbsys.2018.01.010
Received 12 April 2017; Received in revised form 15 October 2017; Accepted 20 January 2018

geocoding (Roongpiboonsopit & Karimi, 2010). To obtain the complete and accurate industrial category values, and the multi-scale text address and coordinates for each enterprise, imputation is required when filling missing values and information (Luengo, García, & Herrera, 2012).

Imputation however, introduces troublesome computing challenges when data volume is big. Enterprise registration data is in a short text format and text-based data imputation involves Natural Language Processing (NLP) techniques, such as short text classification and matching. This process is computing intensive and may result in the Out Of Memory and Intolerable Calculation-Time problems on a stand-alone computer when data volume is big. High Performance Computing (HPC) frameworks are often used to handle the computational issues of big data (Yang, Huang, Li, Liu, & Hu, 2016). Previous research explored big text data processing based on HPC frameworks. However, few studies have systematically investigated HPC-based imputation for big georeferenced text data that involves short text classification, location imputation and geocoding. Moreover, the discussion and applications of such technologies in regional and social science is insufficient in literature.

To fill this gap in the research and solve the big data quality problems endemic to this enterprise registration data, we propose an imputation framework and develop parallel imputation methods based on cutting-edge HPC technologies, to make this data more applicable. An effective solution to these kinds of data quality problems is relevant in many other domains where the use of big data is impeded by incompleteness and the ambiguity issues, especially for big georeferenced text data classification and location geocoding. We compare several widely used text classification methods employing NLP based on Apache Spark to fill missing industrial category values, in terms of accuracy, execution time, memory consumption, and scalability. We also introduce a location imputation method to fill the missing location information and obtain coordinates of each enterprise. Using these imputation results, we briefly analyze the spatiotemporal distribution of all industries in China at multiple spatial scales to illustrate potential applications of this data for analysis of urban spatial structures, urban agglomerations, industrial aggregations, and socioeconomic activities.

This article is organized as follows. Section 2 reviews relevant research. Section 3 introduces the data and HPC-based imputation framework. Section 4 describes industrial category and location imputation. Section 5 details the data imputation experiments and briefly analyzes the potential applications of generated data. Section 6 concludes this article and discusses future research.

## 2. Related work

### 2.1. Industrial spatial distribution analysis

Analysis of industrial spatial distribution has been highlighted in economic geography, urban spatial structure, and regional policy studies. Substantial studies on economics have analyzed the geographical concentration of industries, and the effects of agglomeration economies (Combes, Duranton, Gobillon, & Roux, 2008; Puga, 2010); by analyzing the industrial spatial distribution, many scholars have tried to explain the urban spatial structures (Giuliano & Small, 1991; Liu & Wang, 2016), improve the land use efficiency (Huang, He, & Zhu, 2017), as well as reveal the impact of regional policy on economic activities (Li et al., 2015). In these studies, enterprise or firm data is widely used, including aggregated data, and micro enterprise data. As distinct from aggregated data, micro enterprise data allows users to analyze information at varying spatial levels or partitions, and provides much more fine-grained individual information, offering the potential for theoretical innovation in economic geography and regional studies that are invisible in aggregated data sets (Domenech, Lazzeretti, Molina, & Ruiz, 2011; Lennert, 2011).

The use of micro enterprise data is a promising path in studies of industrial spatial distributions. In economic geography, based on micro enterprise data, distance based spatial agglomeration or clustering of enterprises (Duranton & Overman, 2005; Marcon & Puech, 2010), enterprise heterogeneity (Bernard, Jensen, Redding, & Schott, 2011), and continuous spacial modeling of enterprises (Arbia, 2010) are an active area of research. Domenech et al. (2011) conclude that the use of microdata allows for much richer and detailed results than classical approaches with aggregated data. Using micro data, studies showed how economic activities were shaped by government and market forces (Li et al., 2015), and urban planning suggestions were proposed to optimize the urban spatial structure for achieving greater efficiency (Zhu & Chen, 2007). Many insightful case studies on urban spatial structure and urban expansion have been completed at the inter-city, regional (Liu, Derudder, & Wu, 2016), city and intra-city scales (Gao, Huang, He, Sun, & Zhang, 2016). Further studies are required to examine subcenters within urban districts (Liu & Wang, 2016). In addition, spatiotemporal distribution analysis at multiple scales demands comparison of different regions and industries, as spatial distributions of different industries vary over time and across space (Kneebone, 2010). Industrial spatial distribution analysis therefore, needs chronological, micro-level enterprise data with reliable, accurate, and complete industrial category and location information.

Currently, there have been insufficient studies of the spatial distribution of industries from a multi-scale and temporally sensitive perspective incorporating all kinds of enterprises. It is probably due to the lack of complete data for enterprises containing industrial operating periods, industrial categories and precise location information in the form of geographical coordinates. Big enterprise registration data collected from multiple regional AIC bureaus of China will enable and support such analysis; however, effective use of this data is impeded by data quality problems. To fill this gap in the research, we propose a HPC-based imputation framework to solve the big data quality problem endemic to enterprise registration data, making this data more readily applicable to researchers, planners and decision-makers.

### 2.2. Data imputation in big data era

Imputation has been widely used to fill the missing values in various data types. Datasets, including numerical data and text data, are prone to missing value problems. For numerical data imputation, mathematical methods can be directly used to provide the estimated values for incomplete data by analyzing the relation between multiple data fields or the relationship between different records in the same data field (Luengo et al., 2012; Sim, Kwon, & Lee, 2016). Different from numerical data imputation, text data imputation can harness NLP for semantic analysis. For text data imputation, to label a text with predefined categories, text classification is required; to extract unambiguous location information and even accurate coordinates from georeferenced text data, location estimation and geocoding are often needed (Chen, David, & Yang, 2013; Lennert, 2011). For example, there is a need to classify, estimate and geocode text location for social media data (Barapatre, Meena, & Ibrahim, 2016; Ghahremanlou, Sherchan, & Thom, 2015; Krumm & Horvitz, 2015). Classification, location estimation and geocoding are quite important to georeferenced text data processing.

Short texts are more intractable to be processed than normal document. Short texts are much shorter, nosier, and sparser. For example, a tweet has at most 140 characters (Sun, 2012) and it does not provide sufficient word occurrence, thus impeding traditional text representation methods for classification, such as “bag of words” model (Sriram, Fuhry, Demir, Ferhatosmanoglu, & Demirbas, 2010). For short text category imputation, most existing approaches try to enrich the representation of a short text using additional semantics derived from external sources such as Wikipedia and WordNet (Hu, Sun, Zhang, & Chua, 2009). These methods are limited by the completeness of external corpus, and the lack of semantic-consistency between external corpus and the classified short texts (Zhang & Zhong, 2016), especially in domain-specific studies (Wu, Morstatter, & Liu, 2016). The accuracy of