Health & Place 36 (2015) 35–46

Contents lists available at ScienceDirect

Health & Place

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

A common spatial factor analysis model for measured neighborhood-level characteristics: The Multi-Ethnic Study of Atherosclerosis

Rachel C. Nethery[a,*], Joshua L. Warren[b], Amy H. Herring[a], Kari A.B. Moore[c], Kelly R. Evenson[d], Ana V. Diez-Roux[e]

[a] University of North Carolina at Chapel Hill, Department of Biostatistics, Gillings School of Global Public Health, Hall, CB # 7420, Chapel USA
[b] Yale University, Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA
[c] Drexel University, Department of Epidemiology and Biostatistics, Philadelphia, PA, USA
[d] University of North Carolina at Chapel Hill, Department of Epidemiology, Gillings School of Global Public Health, Chapel Hill, NC, USA
[e] Drexel University, School of Public Health, Philadelphia, PA, USA

ARTICLE INFO

Article history:
Received 12 January 2015
Received in revised form
11 June 2015
Accepted 26 August 2015

Keywords:
Bayesian analysis
Factor analysis
Spatial statistics
Body mass index

ABSTRACT

The purpose of this study was to reduce the dimensionality of a set of neighborhood-level variables collected on participants in the Multi-Ethnic Study of Atherosclerosis (MESA) while appropriately accounting for the spatial structure of the data. A common spatial factor analysis model in the Bayesian setting was utilized in order to properly characterize dependencies in the data. Results suggest that use of the spatial factor model can result in more precise estimation of factor scores, improved insight into the spatial patterns in the data, and the ability to more accurately assess associations between the neighborhood environment and health outcomes.

1. Introduction

Observational studies typically collect large quantities of detailed information on participants in order to identify risk factors for adverse health related outcomes. Researchers working with these data often encounter the need to reduce the dimensionality, a measure of data size, of a dataset in order to facilitate inference in modeling or to generally better understand the underlying structure of a large set of highly correlated measured variables. Factor analysis, a procedure that identifies and estimates a relatively small number of latent variables that capture variability in a larger set of observed variables, can be used to both reduce dimensionality and explore the data structure (Rowe, 1998).

The growing interest in the effects of spatial context on health and the growing availability of large amounts of spatially referenced data have led to an explosion of spatial variables in observational studies. Because these spatial variables are often interrelated, the development of techniques that allow for the exploration of relationships and reduction of dimensionality in the presence of spatial correlation is critical. In this study, we analyze data from the Multi-Ethnic Study of Atherosclerosis (MESA). MESA collects information on over 45 spatial (neighborhood) variables characterizing built and social environments, and researchers stand to gain great insights from the reduction and summarization of these data. However, in the case of neighborhood environment measures, spatial correlation in the variables may be present, and assumptions of independence may be invalid as a result. When working with such spatially referenced data, a common factor analysis may be inadequate because it neglects the potential spatial dependencies in the responses, resulting in violated model assumptions, incorrect and misleading standard error estimates for key model parameters such as the factor scores, and, as a result, the potential for incorrect inference (Rowe, 1998). Spatial factor models are needed to circumvent these issues.

Previous applications of spatial factor models have varied considerably in their methods and purposes, including both multiple and single latent factors, utilizing both continuous and discrete outcome data, adding temporal components, and using the models for prediction (Hogan and Tchernis, 2004; Liu et al., 2005; Lopes et al., 2008; Mezzetti, 2012; Stakhovych et al., 2012; Wang and Wall, 2003). See Table 1 of Stakhovych et al. (2012) for a summary of the spatial factor analysis literature. Building on these

[*] Correspondence to:135 Dauer Drive, 3101 McGavran-Greenberg Hall, CB # 7420, Chapel Hill, NC 27599-7420
E-mail address: nethery@live.unc.edu (R.C. Nethery).

http://dx.doi.org/10.1016/j.healthplace.2015.08.009
1353-8292/© 2015 Elsevier Ltd. All rights reserved.

prior studies, our analysis combines and applies elements of Bayesian spatial factor analysis methodology in order to properly analyze the neighborhood measurements from the MESA study. While Lopes et al. (2008) used their model to predict the outcome variables at new time points and unobserved spatial locations and Wang and Wall (2003) predicted the values of the latent factors at previously observed locations, to our knowledge, ours is the first study to predict the values of the latent factors at unobserved spatial locations. Furthermore, our model is the first to assume a common spatial structure for each of the latent factors without the inclusion of an independent source of variability unique to each location, referred to as the nugget effect in spatial modeling. Finally, our model is introduced and implemented in the point-referenced spatial data setting. Previous studies have introduced these point-referenced models but often work with areal data in the application (Hogan and Tchernis, 2004; Wang and Wall, 2003).

Our analysis begins with the standard non-spatial Bayesian factor analysis model to reduce the dimensionality of a set of MESA neighborhood environment variables. In order to take into account the presumed correlation between the factor scores based on spatial proximity, a second factor analysis model is implemented that allows for the possibility of spatial correlation between the factor scores. The two models are compared to determine if considerable correlation across space exists in the factor scores and to decide whether the added complexity of the spatial model improves the model fit and interpretation of the factors. Because the goal of factor analysis is often the reduction of data to be used as covariates in a health outcome model, an analysis is presented to compare the precision and accuracy of the results of two regression models using body mass index (BMI) as the outcome and the spatially and non-spatially correlated factors, respectively, as covariates.

Working in the Bayesian setting offers a flexible framework for introducing correlation between the latent factor scores. Bayesian estimation is implemented using Markov chain Monte Carlo (MCMC) sampling algorithms which provide samples from the posterior distribution of the parameters. An analysis of these posterior distributions, when correlations in the data are appropriately accounted for, results in correct characterizations of uncertainties in parameter estimates. Bayesian factor models have been previously discussed in the literature (Ghosh and Dunson, 2009; Lopes and West, 2004; West, 2003), and Rowe (1998) provides a comparison between frequentist and Bayesian versions of the factor model.

Given the importance of spatially referenced data in the health research community, our analysis has the potential to lend insight to a multitude of other analyses and research projects. In general, an expanded understanding of the spatial nature of a set of measurements, which can be achieved by applying this methodology, will lead to more accurate analyses, due to improved parameter estimation and correct standard errors for the factor score estimates. This model also allows for the prediction of factor scores at new locations, without the need to collect the full set of original covariates at these new locations. For researchers using cohort data, this ability to predict will be useful when participants move during follow-up. In addition, researchers with an interest in associations between neighborhood environment and health outcomes will benefit from the ability to properly reduce neighborhood data dimensionality, potentially enabling more efficient computation and more concise inference in assessing such associations without substantial loss of information.

2. Materials and methods

2.1. Data description

MESA is an ongoing population-based, longitudinal study designed to explore subclinical cardiovascular disease prevalence and progression in the United States (US), as well as to investigate its association with other health and lifestyle factors (Bild et al., 2002). Approval for MESA participant enrollment and data collection was obtained from the Institutional Review Board at each study site and the coordinating center. From 2000 to 2002, study sites in six US cities recruited 6814 men and women, aged 45–84 years. The sample is 38% white, 28% African American, 23% Hispanic, and 11% Asian. Participants completed questionnaires and participated in a physical examination. For participants in the MESA Neighborhood Study, researchers geocoded the latitude and longitude of each participant's home residence and collected information about the surrounding neighborhood, such as the density of many varieties of restaurants and stores and the crime rates within buffers of various sizes, centered at the residence and workplace. In total, participant information has been collected at five clinic exams as well as through a number of follow-up phone calls (Bild et al., 2002; MESA Coordinating Center, 2014).

The presented analyses utilize data from the Chicago study site ($n=1161$) at Exam 2 ($n=1053$), which occurred between July 2002 and February 2004, and the analyses are restricted to participants who completed Exam 2 in 2003 ($n=815$). Chicago was selected due to the availability of crime data while Exam 2 is chosen to maximize the sample size for a single year. In order to attain the necessary spatial accuracy, only data from locations that are geocoded at the street or zip+4 levels are included ($n=804$). Furthermore, locations are included in the analysis only if their one-mile buffers are contained entirely within the Chicago city limits ($n=603$). Participants with the same spatial coordinates (which indicate participants living in the same house or building) have the same neighborhood measurements. Given that our interest lies exclusively in these neighborhood measurements, only the unique locations are included in the spatial analysis ($n=376$). An additional participant was removed due to inconsistent spatial information, resulting in a final sample of 375 unique locations across Chicago and all with complete data for each of the measurements included in the analysis. The study includes participants that moved within Chicago between baseline and Exam 2, providing greater spatial coverage across Chicago than was originally present in the baseline sample.

In a factor analysis, a fixed set of variables is compiled at the outset to be the subject of reduction and summarization. The following 21 mutually exclusive buffer level variables are included in the presented factor analysis: the kernel density of grocers, supermarket chains, supermarket non-chains, deli/meat/fish/dairy stand-alone stores, liquor stores, drinking places (alcohol), fast food chains, fast food non-chains, other eating places, and total recreational facilities, as well as the percent of land devoted to residential use, the percent of land devoted to commercial use, population density per square kilometer (km), yearly average outdoor murders (per 1000 persons), yearly average indoor murders (per 1000 persons), yearly average outdoor criminal offenses (per 1000 persons), yearly average indoor criminal offenses (per 1000 persons), yearly average outdoor incivilities (per 1000 persons), yearly average indoor incivilities (per 1000 persons), yearly average outdoor assault and battery (per 1000 persons), and yearly average indoor assault and battery (per 1000 persons). A buffer level of one mile is chosen for data completeness purposes and because it represents a common choice in past MESA analyses (Moore et al., 2008, 2009). Numeric summaries of these included variables are displayed in Table 1. In Fig. 1 of the Online