Chapter 3

A Comparison of Principal Component–Based and Multivariate Regression of Cardiac Disease

Fox Underwood
University of Calgary, Canada

Stefania Bertazzon
University of Calgary, Canada

ABSTRACT

Selecting factors suitable to use in a regression model is often a complicated process: the researcher strives to retain all theoretically important factors while avoiding high correlations among independent variables. This chapter models cardiac disease and compares the explanatory ability of component-based multivariate regression models, created through the use of principal component analysis (PCA), with that of direct variable-based, multivariate regression models. The variable-based demographic and socio-economic model contains education, sex, and 3 age factors; in contrast, the component-based model contains age as well as several modifiable risk factors: education, income, family, and housing factors. Moreover, the latter model also has statistically higher explanatory power. Components made through data reduction techniques may not always be interpretable, but, given closer examination of individual components, a component-based model becomes more interpretable. Further, all important factors will potentially be present in models. As such, component-based modelling can be a useful tool for research and public health planning. A key limitation of this work, to be addressed in future research, is the use of a variable (cardiac catheterisation procedures) that remains a crude proxy for cardiovascular disease. More effective analysis will be performed as data becomes available. Exploration into the relationship of factor and their spatial patterns will also be considered.

INTRODUCTION

Many primary health concerns of contemporary societies are inherently spatial: effective accessibility to health care services, prompt and efficient response to epidemic outbreaks, detection and monitoring of environmental health hazards, and consequent urban and regional planning. Management and planning decisions are often supported by quantitative models; indeed the use of quantitative and statistical methods in health research is well established, and the application of spatial analytical methods has gained acceptance over the last several years (Elliott et al., 2000; Waller and Gotway, 2003; Elliott and Wartenberg, 2004). It has been argued that the integration of analytical and visual methods can improve the effectiveness of spatial analysis as a decision support tool for policy and management (Guo, 2007). Further, it has been argued that, though some difficulties still exist, geographic information science (GIS) has a potential role in improving public health (Rushton, 2000).

Despite this potential in assisting management and planning decisions, the use of quantitative methods to spatial data remains a delicate process, and its simplistic application does not necessarily yield reliable and relevant results. One well known problem is that of estimating uncertainty, which stems from two intrinsic properties of geographical phenomena: spatial dependence (i.e. near things are more related than distant things) and spatial non-stationarity (i.e. inconstant variability over space) (Cliff and Ord, 1981). Violation of either assumption inflates the variance—and hence the uncertainty—of the regression estimates, resulting in less reliable models (Anselin, 1988). An additional source of model instability is cross-correlation of the explanatory variables, known as multicollinearity in the model (McGarigal et al., 2001). A common strategy to avoid multicollinearity or near-multicollinearity is a drastic model selection, where only explanatory variables that display low cross-correlation are retained in the model (McGarigal et al., 2001).

This paper presents the use of multivariate regression analysis to identify demographic and socio-economic variables significantly associated with cardiovascular disease prevalence in a Calgary: a large Canadian city. In consideration of the available data (i.e. spatially aggregated records) spatial regression techniques are applied in the presence of spatial dependence to enhance the reliability of the model parameters. In response to multicollinearity, the use of principal component analysis (PCA) (McGarigal et al., 2001) is proposed, leading to the comparison of regression models which make use of orthogonal components and of selected uncorrelated variables, respectively.

Selecting factors suitable to use in a regression model is often a complicated process: the researcher strives to retain all theoretically important factors while avoiding high correlations among independent variables. This tends to result in a compromise rather than a success as not all important factors can always be retained. 'Rules of thumb' and intuition are also often used when deciding on factors to retain; the larger the pool of factors, the more time-consuming this process is. Such effort can be reduced by employing sets of components gleaned through data reduction techniques rather than sets of variables.

BACKGROUND

Study Area and Data

Cardiovascular diseases are among the leading causes of death in the developed world (Kaplan and Keil, 1993; Manuel et al., 2003; Gerber et al. 2006; Canadian Heart Health Strategy and Action Plan, 2009). Disease occurrence is related to personal characteristics such as age, sex, genetic background, and the simultaneous presence of other conditions; all these are known as non-modifiable risk factors. In addition to these, disease prevalence has been found in association with a number of modifiable risk factors,

16 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/comparison-principal-component-based-multivariate/69047

Related Content

To Emphasize Openness
Ken Hartness (2012). *International Journal of Applied Geospatial Research (pp. 1-5).*
www.irma-international.org/article/emphasize-openness/65555

The Curriculum Development of a BIM Resilience Program for the National Institute of Building Science Facility Module
Alan Redmond, Bob Smithand Deke Smith (2014). *International Journal of 3-D Information Modeling (pp. 49-60).*
www.irma-international.org/article/the-curriculum-development-of-a-bim-resilience-program-for-the-national-institute-of-building-science-facility-module/106852

Baseline Climate Grid Resolution and Climate Time Step Impacts on Desert Vegetation Habitat Models
Ross J. Guidaand Scott R. Abella (2020). *International Journal of Applied Geospatial Research (pp. 79-100).*
www.irma-international.org/article/baseline-climate-grid-resolution-and-climate-time-step-impacts-on-desert-vegetation-habitat-models/262167

sUAS Multispectral Survey of the Historical Landscape of Chateau de Balleroy, Normandy, France
Jon W. Carroll (2020). *International Journal of Applied Geospatial Research (pp. 64-78).*
www.irma-international.org/article/suas-multispectral-survey-of-the-historical-landscape-of-chateau-de-balleroy-normandy-france/262166

Agriculture and Conservation in the Natura 2000 Network: A Sustainable Development Approach of the European Union
Cristian Ioja, Laurentiu Rozylowicz, Maria Patroescu, Mihai Nitaand Diana Onose (2013). *Geographic Information Systems: Concepts, Methodologies, Tools, and Applications (pp. 1276-1296).*
www.irma-international.org/chapter/agriculture-conservation-natura-2000-network/70505