Study Profile

Application of Gaussian Process Regression (GPR) in Estimating Under-five Mortality Levels and Trends in Iran 1990 – 2013, Study Protocol

Parinaz Mehdipour BSc[1,2], Iman Navidi BSc[1,2], Mahboubeh Parsaeian MSc PhD Candidate[1,2], Younes Mohammadi MSc PhD Candidate[1,2], Maziar Moradi Lakeh MD[3], Ehsan Rezaei Darzi BSc[1,2], Keramat Nourijelyani PhD•[1], Farshad Farzadfar MD MPH DSc•[2,4]

Abstract

Background: Searching for the latest methods of estimating mortality rates is a major concern for researchers who are working in burden of diseases. Child mortality is an important indicator for assessing population health care services in a country. The National and Subnational Burden of Diseases, Injuries, and Risk Factors (NASBOD) is conducted in Iran with comparative methods and definitions of Global Burden of Disease (GBD) 2010 to estimate major population health measures including child mortality rate. The need to have accurate and valid estimation of under-5 mortality rate led to apply more powerful and reliable methods.

Method: The available datasets consist of under-five mortality rates from different sources including death registration systems and summary birth history (SBH) questions from censuses and Demographic Health Survey. These datasets are gathered at national and subnational levels. We have five time series of under-five mortality rates from SBH method that each one contains 25-year time period. We also calculated Child mortality rates from death registration for 5 years. The main challenge is how to combine and integrate these different time series and how to produce unified estimates of child mortality rates during the course of study. By synthesizing the result of other models, Gaussian Process Regression (GPR) is used as the final stage for generating yearly child mortality rates in this study. GPR is a Bayesian technique that uses data information and defines several hierarchical prior parameters for model. In corporation of GPR and MCMC methods, predicted rates are updated using data and defined parameters in model. This method, also captures both sampling and non-sampling errors and provides uncertainty intervals. The existence of uncertainty for predicting mortality rate is one of the considerable advantages of GPR that distinguish it from other alternative methods.

Discussion: Estimating accurate and reliable child mortality rates at national and sub-national levels is one of the important parts of NASBOD project in Iran. Gaussian Process Regression with its special features improves achievement of this goal. GPR is a serious competitor for other supervised mortality predictive methods. This article aims to explain the application and preferences of GPR method in estimating child mortality rate.

Keywords: Bayesian sampling-based method, child mortality rate, gaussian process regression, NASBOD, study protocol

Cite this article as: Mehdipour P, Navidi I, Parsaeian M, Mohammadi Y, Rezaei Darzi E, Nourijelyani K, et al. Application of gaussian process regression (GPR) in estimating under-five mortality levels and trends in Iran 1990 – 2013, study protocol. *Arch Iran Med*. 2014; **17(3):** 189 – 192.

Introduction

Accurate estimation of the number of deaths in a country, region or worldwide is a crucial step for assessment of the Burden of Diseases.[1] Public health efforts has a substantial focus on improving mortality and morbidity in children.[2] Millennium Development Goal 4 (MDG 4) calls for a reduction of two-third in the under-five mortality rate between 1990 and 2015.[3,4] Reliable child mortality estimates for countries are critical to the monitoring of their progress toward this important goal. There are several methods for estimating child mortality rates and the disagreement for selecting the appropriate method also exists in International Organizations such as WHO, UNICEF and Unit-ed Nation Population Division (UNPD).[1–3]

In order to achieve MDG4 for reduction in child mortality, Iran also makes efforts. National and Sub-national Burden of Disease (NASBOD) project started by Farzadfar, et al.[5] aims to estimate the burden of diseases and attributable burden of risk factors and injuries from 1990 to 2013 at national and sub-national levels in Iran. For estimating burden of disease, both mortality and morbidity data are needed. Due to incompleteness and misclassification of death registration in Iran, [6,7] the project has several steps for mortality estimation. As a key step in this project, we compute levels and trends of under-five mortality rates at national and sub national levels that are also necessary for estimating adult mortality rates. In calculating under-five mortality rates, we face with different data sources at national and sub-national levels over the

Authors' affiliations: [1]Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran, [2]Non-Communicable Disease Research Center, Endocrinology and Metabolism Population Science Institute, Tehran University of Medical Sciences, Tehran, Iran, [3]Department of Community Medicine, Iran University of Medical Sciences, Tehran, Iran, [4]Endocrinology and Metabolism Research Center, Endocrinology and Metabolism Institute, Tehran University of Medical Sciences, Tehran, Iran.
•Corresponding author and reprints: Keramat Nourijelyani, PhD, Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran. Address: Poursina Avenue, P.O. Box 6446, Tehran 14155, Iran. Tel: 98-21-66495859; E-mail: nourik@tums.ac.ir**.**
Farshad Farzadfar MD DSc, Non-communicable Diseases Research Center, Endocrinology and Metabolism Research Institute, Tehran University of Medical Sciences, Tehran, Iran. Address: 4th floor, No. 4, Ostad Nejatollahi St, Enqelab Ave, Tehran, Iran. Postal Code: 1599666615, Tel/Fax: 98-21-88913543, E-mail: f-farzadfar@tums.ac.ir.
Accepted for publication: 10 February 2014

study period. We want to synthesize these data sources with flexible models to provide accurate estimation for the past, present and future of under-five mortality rates at national and sub-national levels. Gaussian Process Regression (GPR) is a statistical technique which is used to synthesize different time trends. For the first time, this technique was employed by Institute of Health Metrics and Evaluation (IHME) to combine data from different sources in estimating child mortality rates.[3]

Statistical techniques provide various methods for combining and integrating multiple trends. Some of these such as Loess, Spline, linear regression models and neural network have a possible advantage in ease of interpretability, but for complex datasets may not, and some of them (like neural networks) are not easy to implement in practice.[8] Most of the models are a linear function of parameters with the uncertainty that is usually expressed as uncertainty of parameters, and do not take into account uncertainty about model structure.[9,10] GPR is a Bayesian method that is more flexible than other alternatives. Unlike classical model fitting approaches, Bayesian algorithms do not attempt to identify the best point estimates of the fitted model. Instead, it computes a posterior distribution over models. These distributions provide a useful way to quantify our uncertainty in model estimates, and to update our knowledge of this uncertainty in order to make more robust predictions on new inputs.[9]

In this article, we present the basic idea on how Gaussian Process models are used to formulate a Bayesian framework for regression and the application of Gaussian Process Regression in under-five mortality estimation.

Materials and Methods

Data Sources

The available datasets consist of under-five mortality rates from six data sources including death registration system from 2006 to 2010, Demographic Health Survey 2000 (DHS 2000) and Iran census for 4 years (1986 – 1996 – 2006 –2011). Under-five mortality rates from Death Registration system are calculated using direct methods. We use indirect methods using summary birth history questions from other data sources. Indirect methods include Maternal Age Cohort (MAC) and Maternal Age Period (MAP) methods which their estimated rates will be combined and smoothed using Loess regression.[3,11–12] The final estimations create, a time series of rates for 25 years prior to each set of dataset.[11] Since Iran administrative divisions has changed and sub-national level during this time span, we computed under-five mortality rates for some of these provinces using reconstruction of population.[13] Finally, our data contains five time series of under-five mortality rates for each province and a few points from death registry. Figure 1 presents an overview of available data sources and their time span. As shown, we want to obtain a unified prediction of under-five mortality rate over the study period and also for future years.

Model Selection

We aim to synthesize these time series and interpolate a unified estimate of under-five mortality rates for a given province-year. The main problem of combining child mortality time series using common methods is failing to embed uncertainty in the final estimations. We can obtain more accurate under-five mortality rate estimations and it's uncertainty by including sampling and non-sampling variances. Gaussian Process Regression is a Bayesian estimation technique that synthesizes mortality rates and interpolates a unique prediction from different data sources. Moreover it can improve previous approaches and captures fluctuation in data caused by sampling and non-sampling errors. Providing smooth trends of child mortality rates over time along uncertainty interval is the achievements of GPR model.

Gaussian Process Model

A Gaussian Process is a generalization of a Gaussian (normal) probability distribution, which extends any finite linear combination of random variables to functions.[13] Like a multivariate normal distribution, Gaussian Process has a mean and a covariance with the difference that these mean and covariance can be functions. A Gaussian Process can completely be described by its mean function and covariance function. The class of Gaussian processes is one of the most widely used families of stochastic processes for modeling dependent data observed over time, or space, or time and space. GP is used as a prior for Bayesian inference. The prior does not depend on data, but specifies some properties of mean and covariance functions.[14–16] Since our target values are under-five mortality rates that are continuous over time, we use Gaussian Process Regression to inferences based on these values.

Gaussian Process Regression for Under Five Mortality

Gaussian Process Regression is used to combine the data, the sampling and non-sampling variance of the data, and other information in terms of model parameters into a single final set of estimates with uncertainty. For each province and source type, the logarithm of under-five mortality rate has a normal distribution with mean and variance. This mean is defined as below:

$$\log_{10} (under - 5mortality\ rates)_{source;\ province} = \\ f\ (year) + \beta_{DRS}\, I_{source=DRS}$$

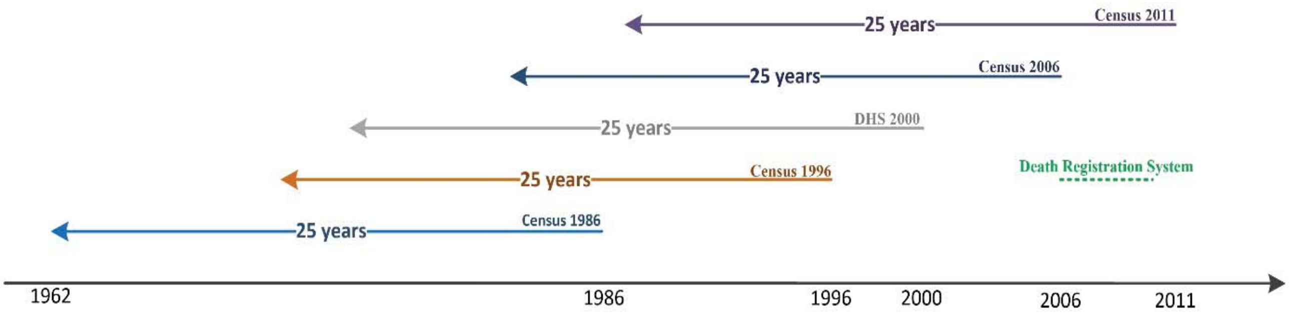

Figure 1. Data sources and their time span

The logarithm of under-five mortality rates modeled as the sum of a function of year (f) and an indicator variable for whether mortality rates are from death registration system (DRS). The logarithm of under-five mortality rate's variance is formed as sampling and non-sampling errors for a given year from each source type. The function f in above equation forms Gaussian Process. The GP has a mean function and covariance function with several parameters which describe the properties of dataset. In the following parts we discuss about these function and their parameters.

Gaussian Process Mean Function

We use Spatial-temporal regression that captures the relationship between child mortality and covariates including years of schooling and wealth index as the GPR mean function.[1] In this framework, observations closer in space and time tend to be more correlated than observations farther away. So, this model can capture any spatial and temporal correlation that remains after accounting for all available covariates. Ignoring such a correlation, as in classic regression models, leads to biased and inefficient estimates.[17] The mean function of a Gaussian Process determines the central tendency of functions from the Gaussian Process distribution.

Gaussian Process Covariance Function

Unlike common regression methods, GPR does not specify a conditional mean function but rather the covariance function between target values. Finding a suitable covariance function of GPR critical part of analysis because it adjusts variation in data by its parameters and data properties. We use "Matern" covariance function,[1,3] which describes the correlation of under-five mortality rates between different years. This function includes three parameters: one is called "amplitude" which calculates deviation between under-five mortality rates and GP mean function over time, another is the "scale" parameter which captures the correlation between years. This parameter is defined by uniform distribution and updated by Bayesian inferences. The third parameter, "degree of differentiability", controls the smoothness of samples from Gaussian Process. Mortality rates obtained from DHS and censuses have disagreement in reporting values and need to make them smoother. We assume less degree of smoothness for death registration observations and greater degree of smoothness for observation from other mortality sources. We have two approaches to choose this parameter: first one is setting the same values that IHME used and maintaining comparability. Another approach is choosing this parameter based on our datasets.

Estimating the completeness of death registration system for under five mortality

The coefficient β_{DRS} estimates the completeness of death registration system. Completeness is defined as the amount of deviation between other data sources and death registration under-five mortality rates. This parameter is modeled as a normal distribution with mean and variance. For estimating their priors, we apply a sub-national level regression of logarithm of mortality rates on time and an indicator variable for adjusting rates whether they come from death registration or not. If the coefficient of indicator variable is statistically significant in regression model, we conclude that death registration system has incompleteness. In this case, we set priors as the mean and variance of indicator variable from regression model. Otherwise, we conclude that death registration system is complete in the province.

Uncertainty with Sampling and Non-sampling Variances

According to existing different data sources for estimating child mortality rates, in addition to considering sampling variance, we are interested in estimating the disagreement between data sources in term of uncertainty. One important output from GPR is an estimate of both the expected value of mortality rates and their uncertainty. The uncertainty estimate captures both uncertainty in prior mean function and the data variance from each observation. Data variance is a function of both sampling and non-sampling error.

We calculate sampling variance of rates assuming a binomial distribution, with sample size equal to the number of children between birth and exact age 5 depending on the data source and probability equal to the under-five mortality rate. The variance that is obtained by using this distribution is sampling error mortality rate.

Non-sampling variance captures the variation between data sources. Unlike the sampling variance, we specify a prior distribution for it. The more variation in data sources, the higher the variance parameter and therefore the more uncertain the measurements. The sum of these two variances will result uncertainty interval.

Markov Chain Monte Carlo

We use Bayesian inference to update predicted mortality rates as a posterior in Bayes rule by combining data and prior probability distribution over parameters in mean, covariance function and the regression model. This predicted series are updated by the data within each province using GPR model. We use Markov Chain Monte Carlo (MCMC) to sample from posterior probability distribution, the distribution describing the probability of a true model given the under-five mortality rates that we observe in a province. Except in special circumstances, posterior distribution for model parameters cannot be calculated analytically and must be approximated. By using MCMC we draw samples from the model's posterior distribution to estimate the median and 95% uncertainty intervals of the mortality rate across country and provinces over years.

To do MCMC calculations, we used RStan package in R software. RStan can employ complex new statistical methods (such as Hamiltonian Monte Carlo) in MCMC. It has similar syntax to BUGS language for Bayesian analysis and compiles C++ codes based on user-written scripts. By using RStan package, we can apply complex models to the data. It is also an open source package which is easily accessible through the internet.

Discussion

In this paper, we described Gaussian Process Regression approach in order to estimate child mortality trends and levels in NASBOD study. One of the major advantages of GPR method is to provide a relevant approach to the assessment of uncertainty, both cross-sectionally and over time, that incorporates sampling error, non-sampling error and parameter uncertainty.[1–3] The uncertainty in this application depends on the sample sizes, non-sampling error for a given data source, the difference between sources and aspects of the Gaussian Process. The GPR method for fitting mortality trends has better out-of-sample performance and predictive validity than the simpler Loess method.[18] Modification of Spatial-temporal as the prior for mean function is one of the strength parts in our model. Because this mean function utilize additional information to take into account how the mortality rates varies across year and province.[17,19]

Gaussian Process gives advantages with respect to the interpretation of model predictions and model selection. It produces easy computations for inference about resulting model.[16] Since the GPR method that is used for mortality prediction is totally an innovative approach for the first time by IHME, it is different and complicated than conventional GPR techniques. This approach specifies several prior for mean and covariance functions and also for their parameters in order to update model prediction by simulation.

IHME has run codes by PyMC package in python, but our analyses are undertaken in R. We employ MCMC to estimate mortality rate during the study period, project future levels of child mortality rate and construct uncertainty intervals for the predicted values. The RStan package is used in this work to provide an efficient way to implement Bayesian approach and draw MCMC simulations.

Despite the above-mentioned advantages, the disadvantages of the GPR approach are the complexity of the fitting procedure and low computational performance due to the application of Bayesian modeling and MCMC simulation. Using multiple priors for functions and parameters slows computational programming speed. It is worth to note that there is not a predetermined package for this method yet and we write all software codes for running GPR model.

GPR models takes advantages of structure in data and use data properties to improve the efficiency of inference for time series data. In fact, GPR is becoming quickly popular for modeling with more robust results.[20]

Author's Contribution

***General designing of paper**: Farshad Farzadfar, Keramat Nourijelyani, Parinaz Mehdipour.*

***Designing of models:** Farshad Farzadfar, Keramat nourijelyani, Parinaz Mehdipour, Iman Navidi, Ehsan Rezaei Darzi.*

***Writing primary draft:** Parinaz Mehdipour, Farshad Farzadfar.*

***Manuscript revision:** Farshad Farzadfar, Mahboubeh Parsaeian, Iman Navidi, Ehsan Rezaei Darzi, Younes Mohammadi.*

Rule of funding Sources

Our research was funded by Iran's Ministry of Health and Medical Education of Islamic Republic of Iran and Setad-e-Ejraie Farmane Imam. The Gaussian Process Regression analysis of these date sets is the MSc thesis of Parinaz Mehdipour in Biostatistics.

Acknowledgments

The authors thank our colleagues in Non Communicable Disease Research Center and appreciate the collaboration of organizations and researchers that provided necessary data for this project. We would also like to express thanks to Dr. Masoud Moradi for his precise editing of the text and Dr. Shirin Djalalinia and Ms Rosa Haghshenas for their efforts on managing coordinative and administrative processes.

References

1. Wang H, Dwyer-Lindgren L, Lofgren KT, Rajaratnam JK, Marcus JR, Levin-Rector A, et al. Age-specific and sex-specific mortality in 187 countries, 1970 – 2010: a systematic analysis for the global burden of disease study 2010**.** *The Lancet*. 2012; **380:** 2071 – 2094.
2. Rajaratnam J, Marcus J, Levin-Rector A, Chalupka A, Wang H, Dwyer L. Worldwide mortality in men and women aged 15-59 years from 1970 to 2010: a systematic analysis. *Lancet*. 2010; **375:** 1704 – 1720.
3. Rajaratnam JK, Marcus JR, Flaxman AD, Wang H, Levin-Rector A, Dwyer L, et al. Neonatal, postneonatal, childhood and under-5 mortality for 187 countries, 1970 – 2010: a systematic analysis of progress towards Millennium Development Goal 4. *The Lancet*. 2010; **375:**1988 – 2008.
4. Murray CJL, Laakso T, Shibuya K, Hill K, Lopez AD. Can we achieve Millennium Development Goal 4? New analysis of country trends and forecasts of under-5 mortality to 2015. *The Lancet*. 2007; **370:** 1040 – 1054.
5. Farzadfar F, Delavari A, Malekzadeh R, Mesdaghinia A, Jamshidi HR, Sayyari A, et al. NASBOD 2013: Designs, Definitions, and Metrics. *Arch Iran Med*. 2014; **17(1):** 7 – 15
6. Farzadfar F, Danaei G, Namdaritabar H, Rajaratnam JK, Marcus JR, Khosravi A, et al. National and subnational mortality effects of metabolic risk factors and smoking in Iran: a comparative risk assessment. *Population Health Metrics.* 2011; **9:** 55.
7. Khosravi A, Taylor R, Naghavi M, Lopez AD. Mortality in the Islamic Republic of Iran, 1964 – 2004. *Bulletin of the World Health Organization.* 2007; **85 (8):** 607 – 614.
8. Rasmussen CE, Williams CKI. Gaussian processes for machine learning. *The MIT Press*, 2006. ISBN 0-262-18253-X.
9. Yizhou F. "Inference for Continuous Stochastic Processes Using Gaussian Process Regression. *University of Waterloo Electronic Theses and Dissertations (UW)*, 2014.
10. Murray-Smith R, Sbarbaro D, Rasmussen CE, Girard A. Adaptive, cautious, predictive control with Gaussian process priors. 2003; 1195 – 1200.
11. Mohammadi Y, Parsaeian M, Farzadfar F, Kasaeian A, Mehdipour P , Sheidaei A, et al. Levels and Trends of Child and Adult Mortality Rates in the Islamic Republic of Iran, 1990 – 2013; Protocol of the NASBOD Study. *Arch Iran Med*. 2014; **17(3):** 176 – 181.
12. Rajaratnam JK, Tran LN, Lopez AD, Murray CJL. Measuring under-five mortality: validation of new low-cost methods. *PLoS Medicine*. 2010; **7(4):** e1000253.
13. Rasmussen C, Williams C. Gaussian Processes for Machine Learning. *MIT Press*, 2006.
14. Kocijan J, Murray-SmithR, Rasmussen CE, Likar B. Predictive control with gaussian process models. *IEEE*. 2003;
15. Neal RM. Monte Carlo implementation of Gaussian process models for Bayesian regression and classication. 1997; Available from: URL: http://arxiv.org/abs/physics/9701026.
16. Andrieu C, De Freitas N, Doucet A, Jordan MI. An introduction to MCMC for machine learning. *Machine learning*; 2003; **50(1-2):** 5 – 43.
17. Parsaeian M, Farzadfar F, Zeraati H, Mahmoudi M, Rahimighazikalayeh G, Navidi I, et al. Application of spatio-temporal model to estimate burden of diseases, injuries and risk factors in Iran 1990 – 2013. *Arch Iran Med*. 2014; **17(1):** 28 – 32.
18. Alkema L, Danzhen Y. Child Mortality Estimation: A Comparison of UN IGME and IHME Estimates of Levels and Trends in Under-Five Mortality Rates and Deaths. *PLoS Med.* 2012; **9:** e1001288.
19. Foreman KJ, Lozano R, Lopez AD, Murray CJ**.** Modeling causes of death: an integrated approach using CODEm. *Population Health Metrics.* 2012; **10**: 1.
20. Rasmussen CE. Evaluation of Gaussian processes and other methods for non-linear regression. *Technical Report*; 1996. Doi:10.1.1.17.729.
21. Surhone LM, Tennoe MT, Henssonow SF. OpenBUGS, 2010.
22. Christensen R Johnson WO, Branscum AJ, Hanson TE. Bayesian ideas and data analysis: An introduction for scientists and statisticians. *CRC Press*, 2011.
23. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. *CRC press*, 2003.