CORMSIS MSc student overview

Witaya Bamrungpong

MSc Data and Decision Analytics
Faculty of Social Sciences
School of Mathematical Sciences

Supervisor: Professor Sujit Sahu

Title: Bayesian spatio-temporal modelling and validation for air pollution data from Piemonte, Italy

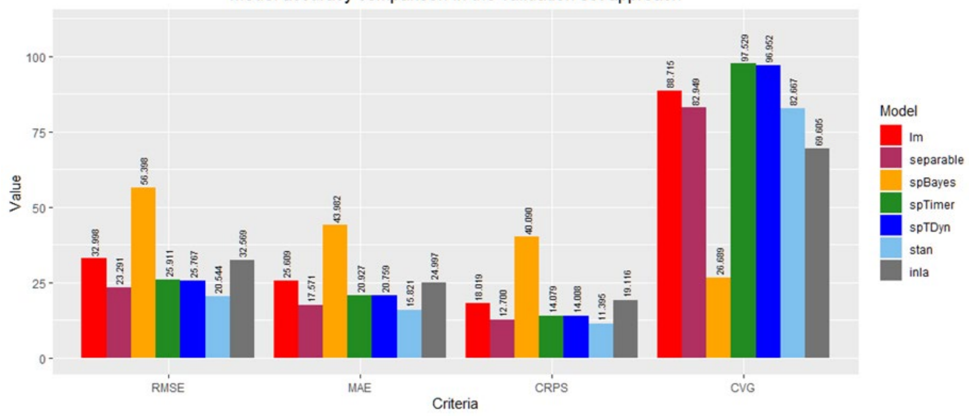
Piemonte, Italy

Data modelling and validation in the age of data science are still challenging tasks in the presence of spatio-temporal correlation. Regression based data science techniques such as linear and ridge-regression do not cope well for such data sets. In this project we experimented with state of the art Bayesian modelling based on Gaussian processes for the purposes of spatio-temporal interpolation of air pollution data. Using a recently developed frontend R-package, *bmstdr*, we fitted several models using well-known R-packages such as *spBayes, spTimer, spTDyn, RStan,* and R-*INLA.* Several validation methods, including the leave-one-out-cross-validation, yielded a space-time interaction model as the best model for spatio-temporal interpolation of a daily observed air pollution data set from Piemonte in Italy during a high pollution time window from October 2005 to March 2006. The validated model was then used to produce maps of out of sample interpolations and their uncertainties.

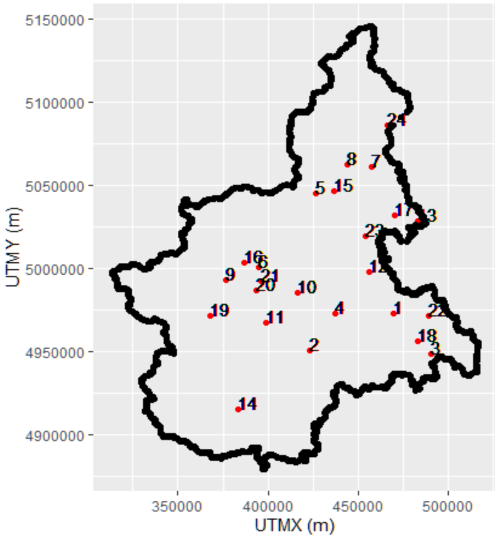
Variable	Description	Unit
pm10	PM_{10} particulate matter concentration	$\mu g/m^3$
s.index	air quality monitoring station	-
UTMX	spatial UTM easting coordinates	m
UTMY	spatial UTM northing coordinates	m
time	number of monitoring day (range from 1 to 182)	-
Date	Date (range from 01/10/2005 to 31/03/2006)	-
Year	year (range from 2005 to 2006)	-
Month	month (range from October to March)	-
Day	day (range from 1 to 31)	-
altitude	altitude	m
ws	daily mean wind speed	m/s
temp	daily mean temperature	K
hmix	daily maximum mixing height	m
prec	daily total precipitation	mm
emi	daily emission rates of primary aerosols	g/s
utmx	spatial UTM easting coordinates	m
utmy	spatial UTM northing coordinates	m

Summary of your MSc project output

Group	Explanatory variable
Spatial data	utmx, utmy
Temporal data	time, Date, Year, Month, Day
Meteorological data	altitude, ws, temp, hmix, prec, emi



Air quality monitoring stations Map

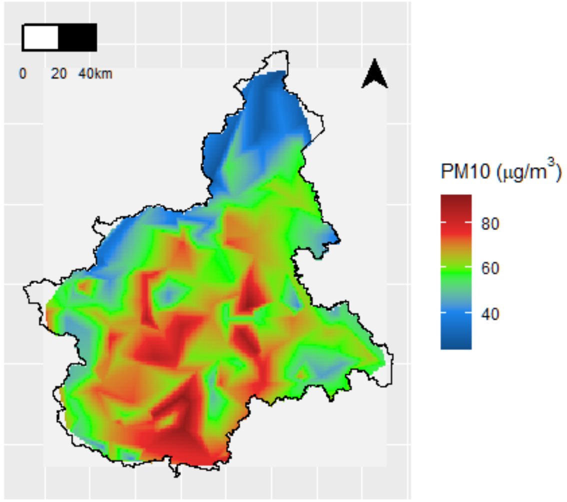


Parameters	Mean	SD	2.5%	97.5%
(Intercept)	7.333	0.096	7.128	7.505
altitude	-0.593	0.032	-0.656	-0.534
ws	-0.193	0.043	-0.274	-0.113
temp	-0.808	0.087	-1.007	-0.651
hmix	0.324	0.065	0.190	0.449
prec	-0.211	0.046	-0.300	-0.120
emi	0.177	0.017	0.144	0.213
utmx	-0.320	0.054	-0.435	-0.222
utmy	-0.400	0.046	-0.498	-0.311
sig2eps	0.193	0.026	0.149	0.257
sig2eta	3.789	0.725	2.755	5.541
phi	0.015	0.003	0.010	0.021

We have found find that a space-time interaction model implemented in the *spTimer* package is the best model according to the standard validation criteria such as root mean square error (RMSE), mean absolute error (MAE), coverage (CVG) and an advanced method called the continuous ranked probability score (CRPS). The middle panel on the left provides a comparison plot of these validation criteria for all the models considered in this dissertation.

Subsequently, we used the best model to spatio-temporally predict PM10 concentration levels in the region for all the days in our study period. These predictions were then used to obtain an average PM10 pollution map and another map (see below) showing the standard deviations of the estimated averages. These predicted pollution map help us to identify the highly polluted areas, which in turn help the authorities in decision making regarding pollution reduction in the region.

Average predicted pollution map of Piemonte between October 2005 and March 2006



Average predicted pollution map of Piemonte between October 2005 and March 2006