**Is spatial machine learning better than the spatial micro-geography OLS model?**

This is a task for passing the ‘Spatial Machine Learning’ class. Below you have data and model framework – the goal is to check if predictions from spatial ML are better than predictions from the OLS model. Please submit R codes, outputs and nice visualisations that solve this challenge. Please consider typical (and/or fancy) machine learning models and their diagnostics (variable importance, partial plots, quality of predictions etc.)

We use a classical dataset on housing valuation in Toledo (U.S.). This raw dataset is available in package {varycoef}. In the link (<https://app.sugarsync.com/iris/wf/D1836703_09297872_6398637>) you get a pre-processed dataset – a few spatial variables were added. The goal of the exercise is to run models on data up to 1997 and test them on the year 1998. The micro-geography model means that it uses information from the local neighbourhood (here, for k=50 nearest neighbours).

**There are two major questions – please answer them:**

**- Are predictions from machine learning models better than those from OLS?**

**- Are ‘spatial variables’ an important explanatory factor?**

Specification of the model (maximum, you can eliminate some variables if needed): The logarithm of the price of real estate depends on its characteristics, location and characteristics in the neighbourhoods.

eq<-log\_price ~ log\_age + log\_lotsize + log\_livearea + Story2more + wall + beds + baths + dGarage + dToledo + MeanPrice + XWX2M + XWX3M + XWX4M + XWX6M + XWX7M + XWX8M + SAFE

**name explanation 🡪 function of variable**

log\_price logarithm of price of real estate (house) 🡪 y variable

log\_age logarithm of the age of real estate 🡪 x - internal features

log\_lotsize logarithm of the size of land (parcel)

log\_livearea logarithm of the surface (floor) of real estate

Story2more dummy to express if there are more than two floors

wall material of construction (partbrk, metlvnyl, stucdrvt, wood, brick, stone, ccbtile)

beds number of bedrooms

baths number of bathrooms

dGarage dummy if there is a garage

dToledo distance to the city centre of Toledo 🡪 x - spatial variables

MeanPrice average of log\_price in the neighbourhoods (k=50 nearest neighbours, knn=50)

XWX2M difference between X and average X of knn=50 for log\_age

XWX3M difference between X and average X of knn=50 for log\_lot\_size

XWX4M difference between X and average X of knn=50 for log\_livearea

XWX6M difference between X and average X of knn=50 for log\_bathrooms

XWX7M difference between X and average X of knn=50 for Story2more

XWX8M difference between X and average X of knn=50 for dToledo

SAFE factor to express in which mini-district the house is located

**# read packages**

library(sf) # for spatial analysis

library(spdep) # for spatial analysis

library(sp) # for spatial analysis

library(varycoef) # original dataset house

library(ggplot2) # for (spatial) ggplot

library(haven) # to read STATA file

library(tidyverse)

# optional packages: to assess the quality of predictions

library(metrica) # <https://adriancorrendo.github.io/metrica/>

library(dplyr)

library(purrr)

library(tidyr)

**# reading Lucas country census track**

# <https://koordinates.com/layer/99834-lucas-county-ohio-census-tracts/>

map<-st\_read("lucas-county-ohio-census-tracts.shp") # NAD83

map<-st\_transform(map, crs=4269) # to NAD83

**# reading data from STATA**

data.file<-read\_dta(file="LucasCountytemp2-NN50.dta") # class tibble

data.file<-as.data.frame(data.file)

**# reading original ‘house’ data from package {varycoef}**

data.pcg<-house # data from package

head(data.pcg)

data.pcg$id<-1:dim(data.pcg)[1]

**# small area fixed effects - SAFE**

**# we get the ID of the region for each observation**

data.sf<-st\_as\_sf(data.pcg, coords=c("long", "lat"), crs=2834)

data.sf<-st\_transform(data.sf, crs=4269) # to NAD83

SAFE<-st\_join(data.sf, map, join=st\_intersects)

data.pcg$SAFE<-as.factor(SAFE$TRACTCE10)

**# merge of datasets – raw and transformed**

data<-merge(data.pcg, data.file, by.x="id", by.y="id")

data.sf<-st\_as\_sf(data, coords=c("long", "lat"), crs=2834)# sf class

data.sf<-st\_transform(data.sf, crs=4269) # convert to NAD83

**# split into train (up to year 1997) and test (year 1998) data**

data.train<-data[data$syear!="1998",]

data.test<-data[data$syear=="1998",]

data.train.sf<-data.sf[data.sf$syear!="1998",]

data.test.sf<-data.sf[data.sf$syear=="1998",]

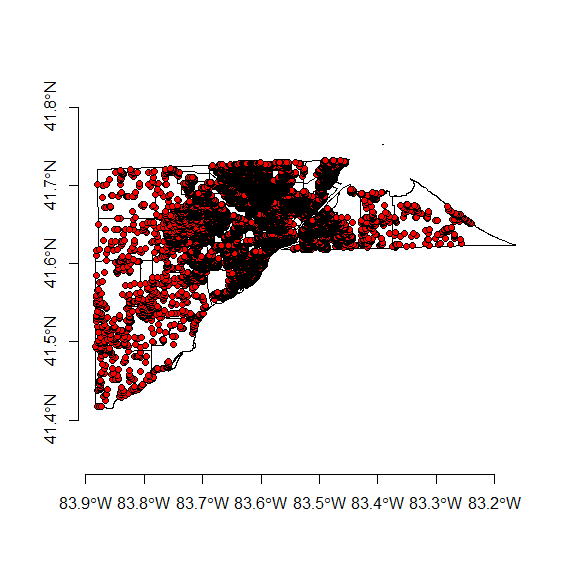
**# plot of data and map**

plot(st\_geometry(map), mar=c(1,1,1,1))

plot(st\_geometry(data.sf), add=TRUE, bg="red", pch=21)

degAxis(1)

degAxis(2)



plot(st\_geometry(map), mar=c(1,1,1,1))

plot(st\_geometry(data.train.sf), add=TRUE, bg="red", pch=21)

degAxis(1)

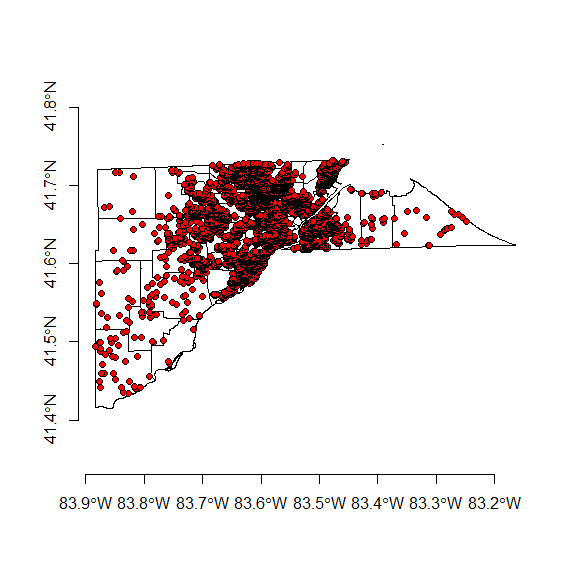
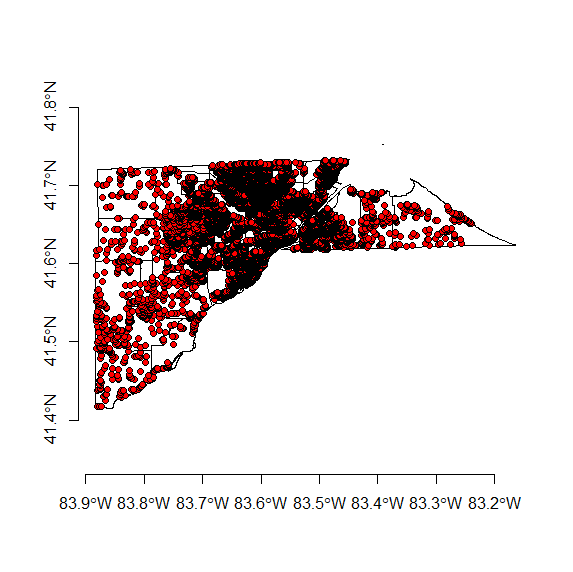
degAxis(2)

plot(st\_geometry(map), mar=c(1,1,1,1))

plot(st\_geometry(data.test.sf), add=TRUE, bg="red", pch=21)

degAxis(1)

degAxis(2)



**# Starter of the model**

eq<-log\_price ~ log\_age + log\_lotsize + log\_livearea + Story2more + wall + beds + baths + dGarage + dToledo + MeanPrice + XWX2M + XWX3M + XWX4M + XWX6M + XWX7M + XWX8M + SAFE

**# OLS model**

model.lm<-lm(eq, data.train)

summary(model.lm)

**# mapping coefficients of SAFE (for small districts)**

map$SAFE.ID<-paste0(rep("SAFE", times=dim(map)[1]), map$TRACTCE10)

coefs<-as.data.frame(model.lm$coefficients)

vec<-rownames(coefs)

coefs2<-data.frame(name=vec, coefs)

colnames(coefs2)<-c("name", "value")

map<-merge(map, coefs2, by.x="SAFE.ID", by.y="name", sort=FALSE, all.x=TRUE)

ggplot(data=map) + geom\_sf(aes(fill=value))+ scale\_fill\_viridis\_c(option="plasma") + labs(title="SAFE – Small Area Fixed Effects")

**# running predictions**

prediction<-predict.lm(model.lm, data.test)

data.pred.OLS<-data.frame(obs=data.test$log\_price, pred=prediction)

missing.OLS<-which(is.na(data.pred.OLS$pred)==TRUE)

data.pred.OLS<-data.pred.OLS[-missing.OLS,] # to eliminate NAs

**And here starts your role – please find a good ML model, better in predictions than this OLS!!! 😊**