Square meter prices and school districts: A case study in spatiotemporal proximity as a tool for prediction and causal inference.

Asger Andersen, ku-id: xzp689, exam number: 137

---

1. Introduction

A common characteristic of big social data is that it is always on [Salganik 2017], which means that much of it is inherently temporal in its nature. If the observations' location are also stored, we often have much more fine grained spatiotemporal data on our hands than what used to be the standard resolution. This kind of new data opens up new ways predicting the behaviour social systems and - more interestingly from a social science perspective - gaining causal insights into the workings of the systems. One the reasons for this is simple that many objects of social science tend to be more alike, the closer they are in space and time. In prediction problems, this means that if one object i is close in space and time to some other objects, we can predict properties of i from properties of the other objects. In causal inference problems, it means that we can use spatiotemporal proximity to control for confounding variables, if objects that are different with regard to the treatment variable still are close in space and time. In this project, I will try to do both, however in quite exploratory ways that definitely leave room for improvements in rigor and thoroughness.

First, I will use spatiotemporal nearest neighbor regression to predict square meter prices of house sales in Copenhagen from 2014-01-01 to 2018-07-31. I will compare the predictive performance of this approach with the predictive performance of a linear regression model with the following predictors: house type (apartment, terrace house and ordinary house), size of the residential area of the sold property, city area, time of the sale, and the percentage of non western immigrants and low income residents in the sale's neighborhood.

Second, I will use spatiotemporal matching to estimate the treatment effect of the Copenhagen schools' grade averages on the square meter prices of the house sales in the school districts. [Gibbons et al. 2013] has done the same, just for the entire of England and with a more sophisticated design, where spatiotemporal matching is combined with a regression discontinuity design. I do not use a regression discontinuity design, which - as far as I understand - is a better design, because it allows the house prices on different sides of a school district border to vary as a smooth function of their distance to the border. Instead I, more restrictively, assume that if house sales on different sides of a school district border are close enough in space and time, they are not *systematically* different in any price-influencing way other than the fact that they belong to different school districts. Again, I compare the resulting estimate of treatment effect with the estimate obtained by a linear regression with

square meter prices as outcome, school grade averages as treatment values, and the following controls: house type, size of the residential area of the sold property, city area, time of the sale, and the percentage of non western immigrants and low income residents in the sale's neighborhood.

Thus, I will try answer two related research questions, namely 1) how well does spatialtemporal nearest neighbor regression predict square meter prices compared to a linear regression model with spatiotemporal, house specific and socioeconomic predictors, and 2) how does the estimate of the causal effect of school districts' grade average on the districts' square meter prices obtained by a spatiotemporal matching approach compare to the estimate obtained by a linear regression model with spatiotemporal, house specific and socioeconomic controls?

2. Data

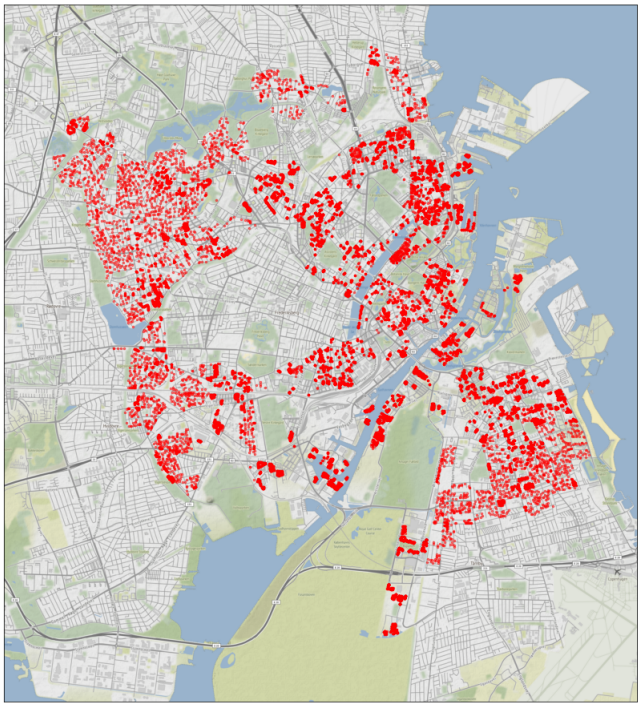
The notebook scraping-and-structuring contains all the code, which implements the scraping and data structuring described in this section. The scraping code is a modified version of code by Snorre Ralund. The original code can be found at github[1].

2.1. House sales data

I obtained data about the house sales by scraping the web page www.boligsiden.dk[2]. I scraped the price, the size of the residential area of the property, the location in latitude/longitude, the city area, the date and the house type (apartment, terraced house or house) of all house sales in the municipality of Copenhagen from 2014-01-01 to 2018-07-31. Here is a map of the 30251 scraped house sales:

[1] https://github.com/abjer/tsds/blob/master/data/scraping_examples/scraping-boligsiden.ipynb

[2] www.boligsiden.dk is a web page, which is owned by all the major real estate agent chains in Denmark. The web page has an archive, which contains data for all the sold houses in Denmark.

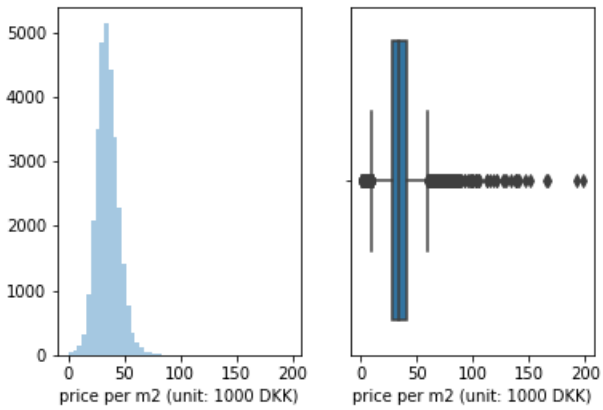


After I had scraped the sales and organized the data in a pandas dataframe, I divided the price by the size of each to get its price per m2. I used the city areas of the sales to define a variable, which denoted whether the sale's location was east or west of the harbour, which splits Copenhagen. I transformed the latitude/longitude coordinates to easting/northing coordinates, which are measured in meters, with the use of the UTM projection, zone 32N[3]. This projection from spherical to planar geometry does not distort distances in any significant way within a small area like Copenhagen.

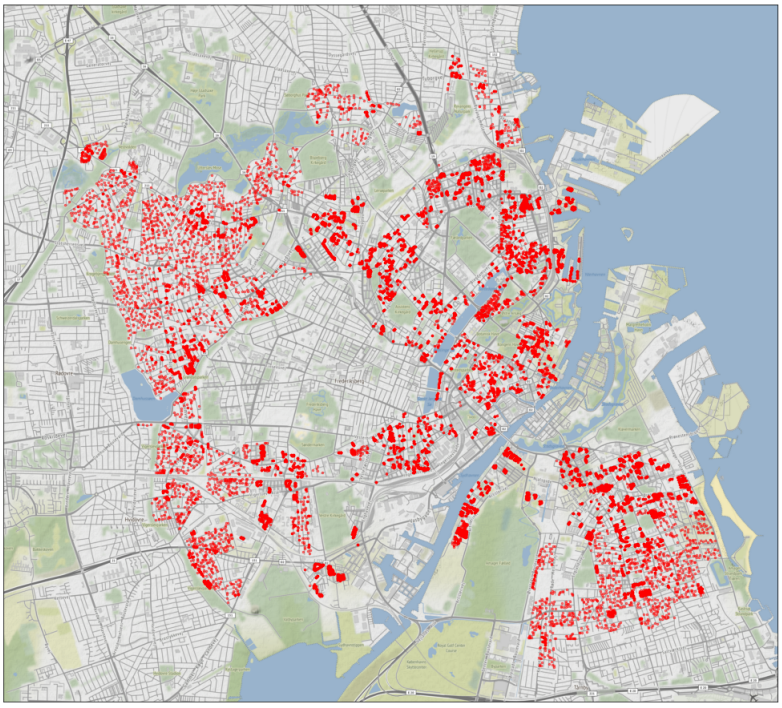
I removed some sales from the data that I was going to use for the causal part of my analysis. First, I removed all sales from four school districts[4], which were too new to have grade data yet. I also removed the sales from the school district Christianhavns Skole, since it is an island and therefore does not border any other school districts in the sense needed by my causal analysis. Also, I removed the sales with square meter prices below the 0.01 and above the 0.99 percentile of the marginal square meter price distribution of the sales. I did this, because the distribution has a lot of heavy outliers:

[3]http://spatialreference.org/ref/epsg/32632/

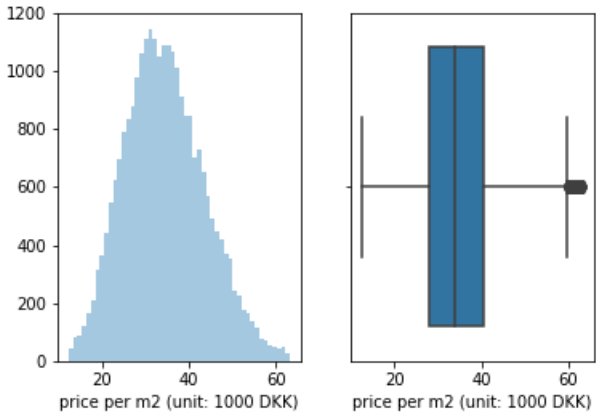
[4]Skolen i Sydhavnen, Kalvebod Flled Skole, Skolen p Strandboulevarden and restad Skole



and I think that whatever made the outliers have extraordinary high or low prices, it is safe to assume that it is not their school districts. Therefore I believe the analysis becomes more robust by excluding them, since they should not be allowed to drive the conclusions of the analysis. All in all, I excluded 4562 of the original 30251 sales. Here is a map of the sales left for the causal analysis:



And here is a plot of the distribution of square meter prices in the filtered data:



2.2. School performance data

For each year between 2012 and 2017, I got each schools grade average from the final exams in the nationally mandatory subjects. I got the data from the open data web page of the Danish ministry of education[5]. For each house sale i I calculated its school grade average sg_i, as the average of the sale's designated school's grade average from the last two years before the year of the sale. I calculated the bordering school average bsg_i of each sale i with the exact same computation, except that I used the grade averages from the nearest neighboring school district of the sale. That means that all sales from the same year within the same school district will have the same school grade average. However, sales from different years can have different school averages, even if they are within the same school district. Two sales will only necessarily have the same bordering school average, if they are from the same year within the same school district and share the same nearest neighboring school district.

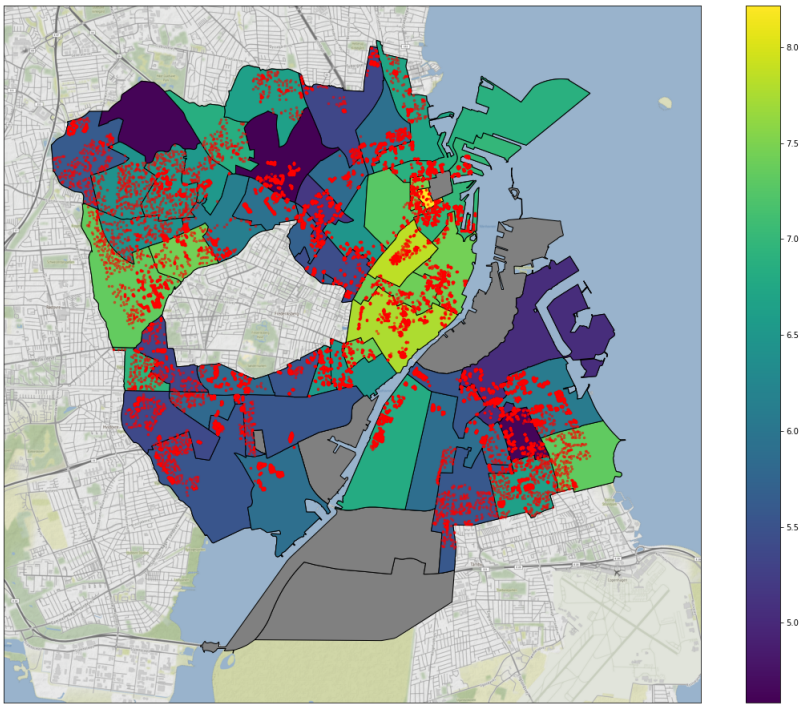
2.3. School district data

I got the location of the school districts from the open data web page of the municipality of Copenhagen[6]. The data is updated yearly, if there have been any changes in the districts. Of course, I should have used each year's school district data for the given year's house sales, and I therefore wrote to the municipality and asked for it. Unfortunately, they did not answer, so I am left with only the most recent edition of the school districts. If there have been many major changes of the school districts from 2014 until now, it would completely destroy the causal part of my analysis, and unfortunately I have not been able to figure out what kind of changes there have been. Therefore, all I can do for now is to assume that if there have been any changes, they have not been big enough to completely mess up my analysis.

As with the location of the house sales, I transformed the geometry of the school districts to planar geometry via the UTM zone 32N projection. After the transformation, I used

[5]https://uddannelsesstatistik.dk/grundskolen/
[6]https://data.kk.dk/dataset/skoledistrikter

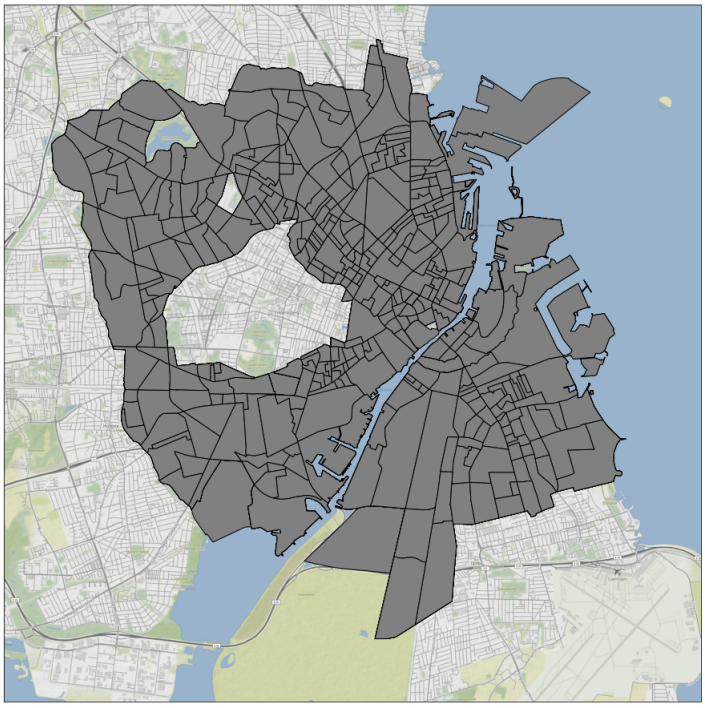
spatial join in Geopandas to merge the school districts on the sales. Here is a plot of the sales (filtered for the causal analysis) and the school districts. The districts are colored by their average grade average from 2012 to 2017 (the grey ones are the ones that I removed from the causal analysis):



2.4. Socioeconomic data

I also got data about the percentage of low income residents and the percentage of non western immigrants in the neighborhoods of Copenhagen from the municipalities open data web page[7]. This data has one observation for each of the so called "roder", which are small neighborhoods that used to be the basic administrative unit for organizing tax collection. As with the location of the house sales, I transformed the geometry of the "roder" to planar geometry via the UTM zone 32N projection. After the transformation, I used spatial join in Geopandas to merge the "roder" with their socioeconomic variables on the sales. Here is a map of how the "roder" are spread across Copenhagen:

[7]https://data.kk.dk/dataset/samlede-socio-data-kbh



3. Analysis

The code that implements the analysis described in this section can be found in the two notebooks called dist-matrix and analysis.

3.1. Spatiotemporal nearest neighbor regression for prediction of square meter prices

In this part of the analysis, I want to see how well spatiotemporal nearest neighbor regression can predict the square meter price of a sale. I therefore randomly split the house sales in two sets: A training set, which constitutes 80 percent of all the observations, and a test set.

Nearest neighbor regression is known as a lazy learner [Raschka 2015], since it does not predict the outcome of the test points by learning any function that maps the test points covariates to their predicted outcomes. Instead it simply uses some metric on the covariate space to determine the nearest neighbors among the training points for each of the test points, where after it predicts the outcome of each test point as the (possibly weighted) average of the outcomes of the test point's nearest neighbors. K nearest neighbors chooses the K training points as nearest neighbors for each test point, where as radius nearest neighbors chooses all the training points within a certain distance of the test point. I use

a version of radius nearest neighbors, where I say that all training house sales within a maximum spatial distance δ and a maximum time difference τ of a test house sale are nearest neighbors of the test house sale. If there are no sales that are appointed as nearest neighbors of a given test sale under these conditions, I appoint the spatially closest training sale to be the test sale's nearest neighbor. I then calculate the test house sale's predicted square meter price as the weighted average of the nearest neighbors' square meter prices, where the weights are set to the inverse of the sum of the distance in space and difference in time. Formally, I predict the square meter price $\hat{y}_i$ of a test house sale i as

$$\hat{y}_i = \frac{1}{\sum_{j \in A_i(\delta,\tau)} w_{ij}} \sum_{j \in A_i(\delta,\tau)} w_{ij} y_j$$

where $A_i(\delta, \tau)$ is the set of training house sales within spatial distance δ and time difference τ of i, and the weights are defined as

$$w_{ij} = \left(\frac{1}{1 + d_{ij} + t_{ij}} \right)^{\alpha}$$

where d_{ij} is the spatial distance and t_{ij} is the time difference between test house sale i and training house sale j. Spatial distances are measured in meters, and time differences are measured in days.

My version of spatiotemporal radius nearest neighbor regression has three hyper parameters, namely the maximum distance δ, the maximum time difference τ and α that controls how fast the weights decay when time and spaces increases between sales. In general machine learning theory, the tuning of hyper parameters is often related to striking an error-minimizing balance between the bias and variance of the algorithms prediction. We can also see this general principle in my predictive algorithm: With small δ and τ the predictions of a test house sale's square meter price will have a large variance under resampling of training house sales, since the prediction will be determined by fewer training house sales. Larger values of δ and τ will mean more nearest neighbors and therefore smaller variance of the prediction under resampling. However, the prediction will also be more systematically wrong (that is, biased), insofar the square meter price of house sales are more alike, when the sales are near in space and time. The inclusions of weights and the hyper parameter α is an attempt to give the algorithm an opportunity to strike a more delicate balance of this bias-variance trade-off.

In order to choose the some good values for the hyper parameters, I use 4 fold cross validation on the training set to estimate the predictive performance measured by by R^2 of each possible combination of $\delta \in \{150, 250, 500\}$, $\tau \in \{365, 730, 1460\}$ and $\alpha \in \{0.5, 1, 2\}$. I find that the optimal combination is $\delta = 150$, $\tau = 730$ and $\alpha = 1$. When I run my algorithm on the test set with this configuration of the hyper parameters, it has an R^2 of 0.52.

When I train a linear regression model with square meter price as outcome and city area[8] and time of sale[9], I get an R^2 of 0.32, when I use it to predict the square meter prices of the

[8]Kbenhavn K, Kbenhavn , Kbenhavn V, Kbenhavn S, Kbenhavn SV, Kbenhavn N, Kbenhavn NV, Valby, Vanlse, Brnshj, Nordhavn, Hellerup and Kastrup.

[9]Each year is divided into the first and last six months.

test house sales. When I also include house specific information in the form of house type[10] and size of the residential area, I get an R^2 of 0.35 on the test set. When I also include the percent of non western immigrants and low income residents of the neighborhoods, I get an R^2 of 0.39 on the test set.

All in all, I can conclude that spatiotemporal nearest neighbor regression predicts square meter prices better than the linear regression models, I have compared it with. This is not surprising, since we would also expect predictors in the regression models to be highly correlated among house sales that are close in space and time, and additionally spatiotemporal proximity also implies similarity in other variables, which are unobserved in the regression models, such as proximity to green areas and shopping opportunities and the social status and atmosphere of the local neighborhood. Although spatiotemporal nearest neighbor regression outperforms the linear regression models predictively, it does not mean, however, that it is a better predictive approach along all dimensions of comparison. For instance, it is much, *much* more computationally expensive, and it does not output a predictive model which has a meaningful interpretation.

3.2. Spatiotemporal matching for causal inference of the effect of school districts on square meter prices

In this part of the analysis, I will - in a quite exploratory and not-at-all-entirely-thought-through way - try to use spatiotemporal matching to estimate the treatment effect of the Copenhagen schools' grade averages on the square meter prices of the house sales in the school districts. Following Rubin's potential outcomes framework [Rubin & Imbens 2015], I will try to conceptualize the values of the treatment variable as assigned interventions, and then ask under what conditions we can consider the assignment of the treatment values as random; or more precisely as independent of the treated units' potential outcomes under all possible treatment values [Rubin & Imbens 2015, Imbens & Hiranu 2004]. In our case, we will consider the value of treatment T_i for a sale i as the difference between i's school grade average sg_i and i's bordering school grade average bsg_i (sg_i and bsg_i were defined for all sales i in the section about school performance data above):

$$T_i = sg_i - bsg_i$$

In general, we cannot consider the treatment value T_i of a treated sale i to have been assigned independently of the sale's potential outcomes under all possible treatment values. This is because we have reason to suspect that sales with high treatment values would have had higher square meter prices on average, even in a potential world where all the schools had the same grade average. We can namely suspect that higher treatment values are correlated with variables such as higher education level of the residents, and that these other variables would cause the sales to have higher average square meter prices, even in a world with no differences between the schools' grade averages. However, within small enough regions in space and time - say within a distance of δ meters and a time difference of τ days - we

[10] Apartment, terrace house or regular house.

can maybe assume that if sales are on different sides of a school district border, then the difference in treatment values are the only systematic, price-influencing difference between the sales on each side of the border. In that case, I think we can unbiasedly estimate the average treatment effect of the treatment value $T = t$ with the following process.

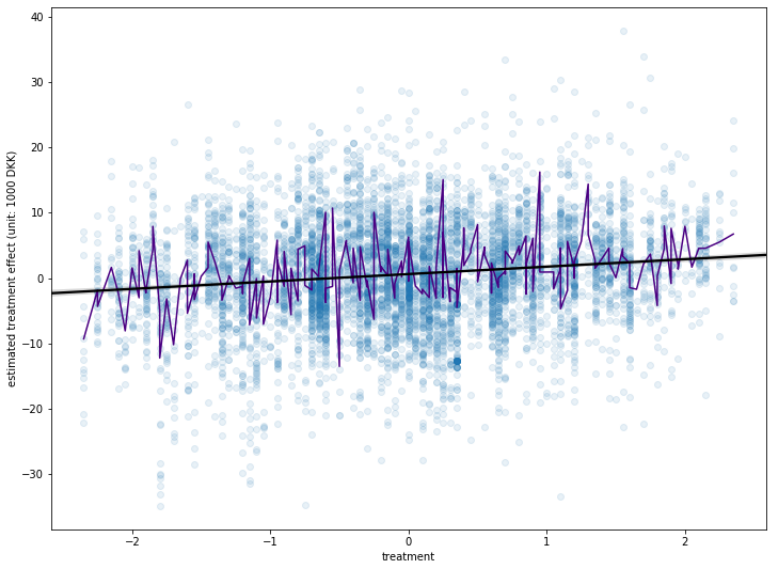
First, we take all sales i with treatment $T_i = t$. For each of these sales i, we then find i's nearest bordering neighbors $B_i(\delta, \tau)$, which are the sales that are located in i's nearest neighboring school district and further are located within δ meters and sold within τ days of i's location and sale date. We then estimate i's counterfactual square meter price $\widehat{cfy_i}$, if i had been located within the nearest neighboring school district instead of being located within its own school district as the average of the square meter price of its nearest bordering neighbors:

$$\widehat{cfy_i} = \frac{1}{|B_i(\delta,\tau)|} \sum_{j \in B_i(\delta,\tau)} y_j$$

We then estimate the treatment effect te_i of treatment T_i on the square meter price of i as the difference between the actual square meter price and the estimated counterfactual one:

$$\widehat{te}_i = y_i - \widehat{cfy_i}$$

It is clear that $\widehat{te}_i$ is a very poor treatment effect estimate for the individual sale i, since i is surely different from the sales in $B_i(\delta, \tau)$ in a lot of other price-influencing ways that just belonging to a different school district. However, I think that the under our assumptions we can get an unbiased estimate of the average treatment effect of $T = t$ by averaging the estimated treatment effects $\widehat{te}_i$ of all sales i with $T_i = t$. Here is a plot of the estimated treatment effects $\widehat{te}_i$ versus the treatment values T_i for all sales i with δ set to 150 meter, and τ set to 730 days:



The purple line is the estimate of average treatment effect for each observed treatment value in our data. We see that this estimate has too much variance to be useful, if we want to model the treatment effect as a smooth function of the treatment value. A simple and interpretable way of smoothing the estimator is simply to assume that the treatment effect is a linear function of the treatment value, and then use linear regression to estimate the effect size. With this approach we get an estimate of a 1110 DKK increase in square meter price per 1 grade increase in the treatment. Interestingly, this is almost the same as the estimate we get from running a linear regression model with the sales' square meter price as outcome, their school grade average as treatment and the following controls: house type, size of the residential area of the sold property, city area, time of the sale, and the percentage of non western immigrants and low income residents in the sale's neighborhood. In this model, the school grade average gets an estimated effect of 1094 DKK increase in square meter price per 1 grade increase.

Unfortunately, I have not had the time to formalize and investigate the rather loose thoughts, which I used to motivate my spatiotemporal matching estimator above, and maybe I am fundamentally mistaken in the way I frame spatiotemporal matching in light of the Rubin potential outcome framework; both areas are new to me, and I have not had the time so far to study the literature thoroughly[11]. It is interesting that the linear regression model and the matching approach produces almost identical estimates of the effect size, but it does not really say of the validity of the matching estimator, unless we have high trust in the validity of the linear regression model. In fact, if we have low trust that this rather simple linear regression model has managed to make a good estimate of the effect size, then the similarity of the estimates should make us suspicious of the validity of the constructed matching estimator.

4. Conclusion

In my first research question, I asked how well does spatial-temporal nearest neighbor regression predict square meter prices compared to a linear regression model with spatiotemporal, house specific and socioeconomic predictors? I conclude that it performs better in terms of purely being able to explain more of the variance in the square meter price. However, this does not mean that it is a better predictor along all dimensions of comparison. In my second research question, I asked how does the estimate of the causal effect of school districts grade average on the districts houseprices obtained by a spatiotemporal matching approach compare to the estimate obtainedby a linear regression model with spatiotemporal, house specific and socioeconomic controls? I conclude that they are almost identical. However, this does not imply very much, unless we have very high or low trust in one of the models in advance. I am personally reluctant so far to commit to high or low trust in either of the models, because I have not had the time to investigate the assumptions of neither of the models, and also because I am in doubt if construction of my matching estimator can withstand more thorough inspection.

[11] I must admit that the reason that I used the Rubin framework was mainly that I wanted to practice working with it, but in retrospect I do not thinking that I have succeeded in applying it well in this case.

5. Litterature

- Salganik, Matthew: 2017: Bit by bit - Social Research in the Digital Age.
- Gibbons, Stephen; Machin, Stephen; Silva, Olmo: 2013: Valuing school quality using boundary discontinuities.
- Raschka, Sebastian: 2015: Python Machine Learning
- Imbens, Guido W.; Hiranu, Keisuko: 2004: The propensity score with continuous treatment.
- Rubin, Donald; Imbens, Guido W.: 2015: Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.