Urbanicity and Mortality Disparities in Andalusia

10.05.2018

General Comment

This document is an attempt to answer the comments on our project on urbanicity and small area characteristics and the effect on health/mortality disparities in Spain and make some comments myself/ask questions. I hope this brings us on the same page, invites further comments, serves us to stay up-to-date, and lead to a quick completion. It will probably be a good idea to have this or a similar document with date stamp in the dropbox so that everybody who gives comments can do so whenever he or she pleases.

Main Idea

Just for repition, the main idea of the project, as far as I understood it, is that there are particular area features which potentially affect health and ultimately mortality of the people who live/have a residence in these areas. Intuitively and backed up by historical research (key concept: *urban penalty*), we would assume that such environmental hazards are concentrated in urban agglomerations (worse water, air quality, heat etc.). As there is no universal approach to even conceptualize what we mean by urban or rural, we decided to create our own measure with satellite land cover data. Based on the comments we have received in Cambridge, it was decided to make a clean separation between purely physical features and socio-environmental factors. The aim of this separation was also to have a re-usable indicator for how urban a place is. Such an indicator can not only be applied for other types of analysis, it would add something different to the demographic research on geographical differences. *Personal note*: Given the comments we have received so far, I would strongly suggest to focus on this "urbanicity" index, even if it does not contribute much to the explanation of mortality disparities in our case. However, we add something by basically decomposing the "small area" effects. Our result is, that is does not matter if an individual resides in a city or in the country (despite arguable advantages of living close to specialized health care centers).

Potential threaths to validity - or just things that are still up for discussion

Population - Age range and time lag between observations

Since we assume individual mortality risks are affected by area features of their residential environments, we excluded age groups with high probability for residential changes (younger people who move out the parents household and older people who move into care facilities or with relatives). The age range is between age 35 and 80. The follow up time is 12 years, thus our oldest individuals are about 92 years old. We can argue that even if they have moved in with younger relatives or in care facilities at higher ages, they were still exposed to the residential environment for a substantial part of their life time.

We discussed the time lag between observations and event/censoring time as threat to validity already and if you agree, I will just mention it in the part on data limitations and try to argue that the variables we have selected are more likely to be time-independent. While that might be true for the individual level variables, I doubt that we see no development in degree of urbanicity or other area features in the last 12 years. If you have ideas on this point, please let me know. The same problem occurs for the residential changes which we cannot observe.

Number of events

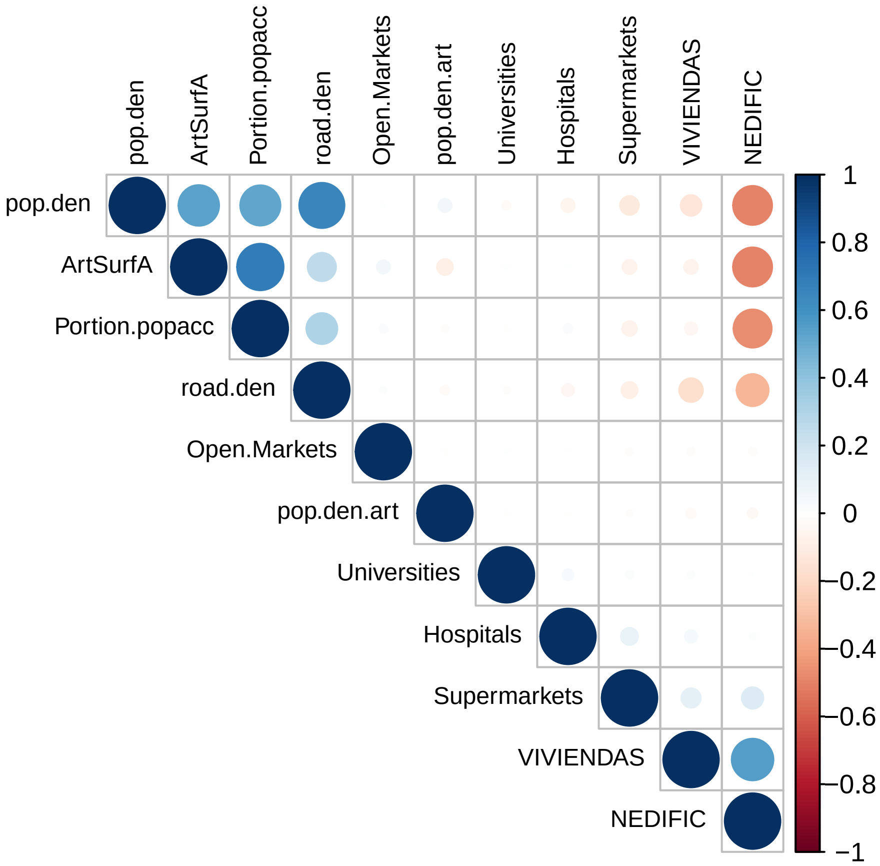
Another potential problem I came across in a paper, can be too few deaths once the individuals are aggregated by census tract. Given the large number of cases and events, it did not seem something to worry about. However, as you can see below, the average number of deaths by census tract is only about NA and there are more than 70 census tracts with less than 2 deaths within the observation period. For now I will not include these tracts and its residents in the analysis, but if you have a better suggestion, please let me know!

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   6.000   9.000   9.354  12.000  33.000
```

```
##  < 2 deaths >= 2 deaths
##          74       5307
```

Urbanicity Index

First of all, I agree with Fran, that the indicator construction is kind of a black box and in the past we had good reasons to be a little bit mysterious about it while we were figuring out the optimal composition and still meeting the deadlines. In order to achieve more clarity I am going to re-write the paragraph on the indicator construction within the next days. As Fran has suggested, it would also be good to show the factor weights or the correlation between the physical variables. Personally, I would prefer to mention that we selected physical variable a,b, and c “apriori” and continued to test correlations and ran a factor analysis. My feeling is that a correlation plot like the following has some descriptive power but it is rather unusual to include it in a paper.



I tried to include just physical and population-based measures, and what we see is not so great but in line with earlier results. We find the four variables artificial surface, population density, road density and proportion of service area (health center) positively correlated. Once they are standardized they seem to make a good indicator which Crohnbachs alpha suggest for expample.

```
##
## Reliability analysis
## Call: alpha(x = CS)
##
##   raw_alpha std.alpha G6(smc) average_r S/N
##        0.9       0.9    0.89      0.69 8.9
##
##  Reliability if an item is dropped:
##              raw_alpha std.alpha G6(smc) average_r  S/N
## POPDEN.I.SD       0.83      0.83    0.77      0.62  4.9
## ARTSURF.I.SD      0.86      0.86    0.85      0.68  6.3
## ROADDEN.I.SD      0.86      0.86    0.82      0.67  6.2
## SERAREA.I.SD      0.92      0.92    0.90      0.79 11.0
##
##  Item statistics
##                 r r.cor r.drop
## POPDEN.I.SD  0.93  0.94   0.88
## ARTSURF.I.SD 0.89  0.83   0.79
## ROADDEN.I.SD 0.89  0.87   0.80
## SERAREA.I.SD 0.79  0.67   0.64
```

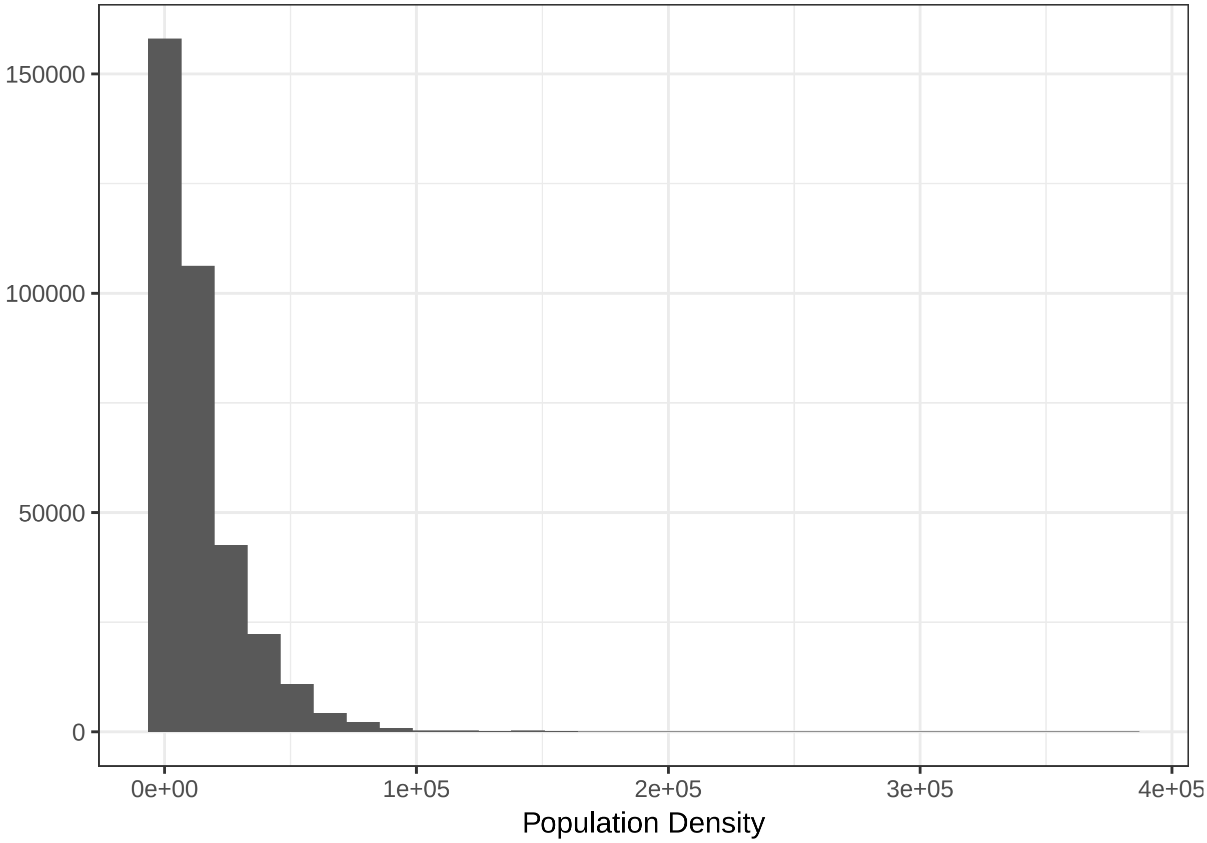
Nevertheless, I agree to make the whole construction less of a black box for the editor/future reader and would suggest to present the factor wheights and elaborate on the selection.

```
##
## Call:
## factanal(x = cor.UI.fac, factors = 1, rotation = "varimax")
##
## Uniquenesses:
##  POPDEN.I.SD ARTSURF.I.SD ROADDEN.I.SD SERAREA.I.SD
##         0.03         0.36         0.20         0.61
##
## Loadings:
## [1] 0.98 0.80 0.89 0.63
##
##                Factor1
## SS loadings       2.79
## Proportion Var    0.70
##
## Test of the hypothesis that 1 factor is sufficient.
## The chi square statistic is 64498.59 on 2 degrees of freedom.
## The p-value is 0
```

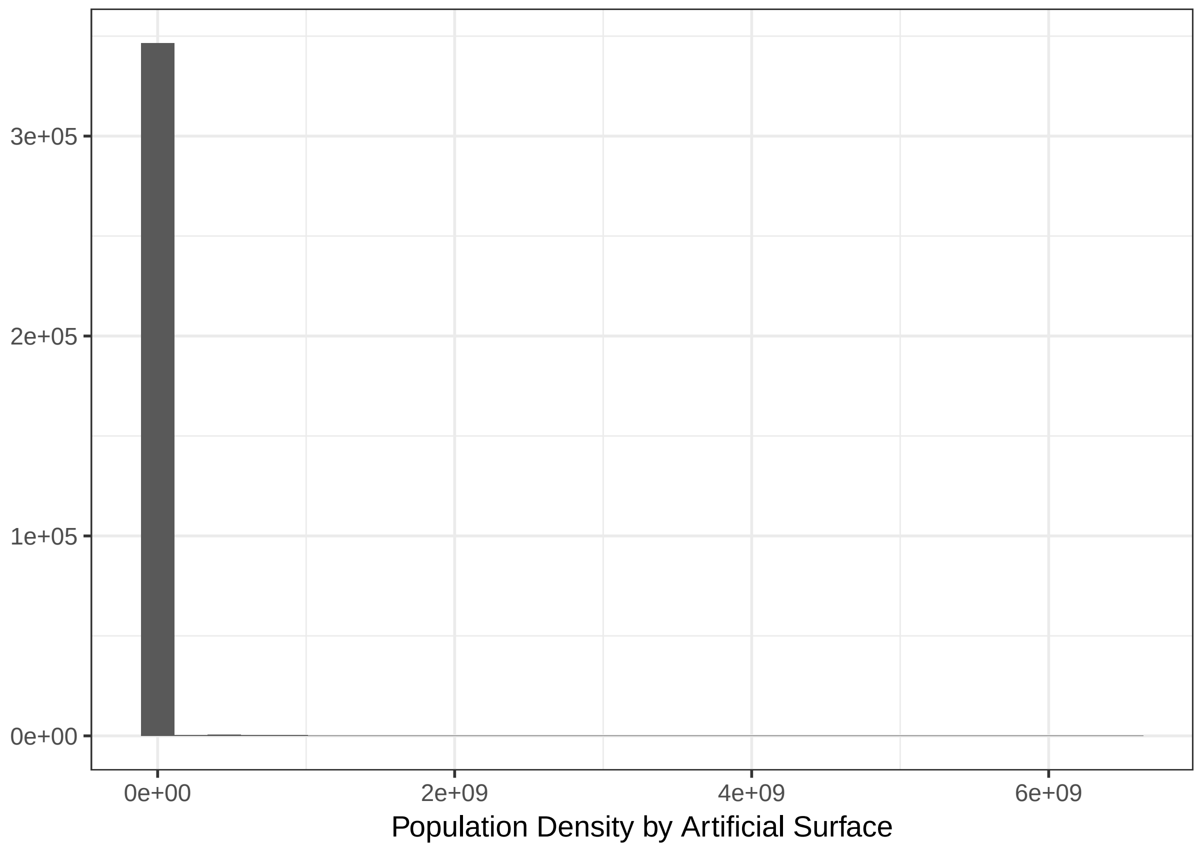
I also completely agree that it will be necessary to shorten the part on the indicator components. This is still an artefact from the rush to the deadline and I am working on it.

Indicator components

Here is room for debate. I would argue that an indirect measure for "urbanicity" needs to include population density. Knowing that most census sections are determined by population size, there is remaining variability in area size. Including population density would therefore add a component of crowdedness which is more or less already determined by the area size of the tract. However, not all tracts of only based on the population in it as we see from the variation.



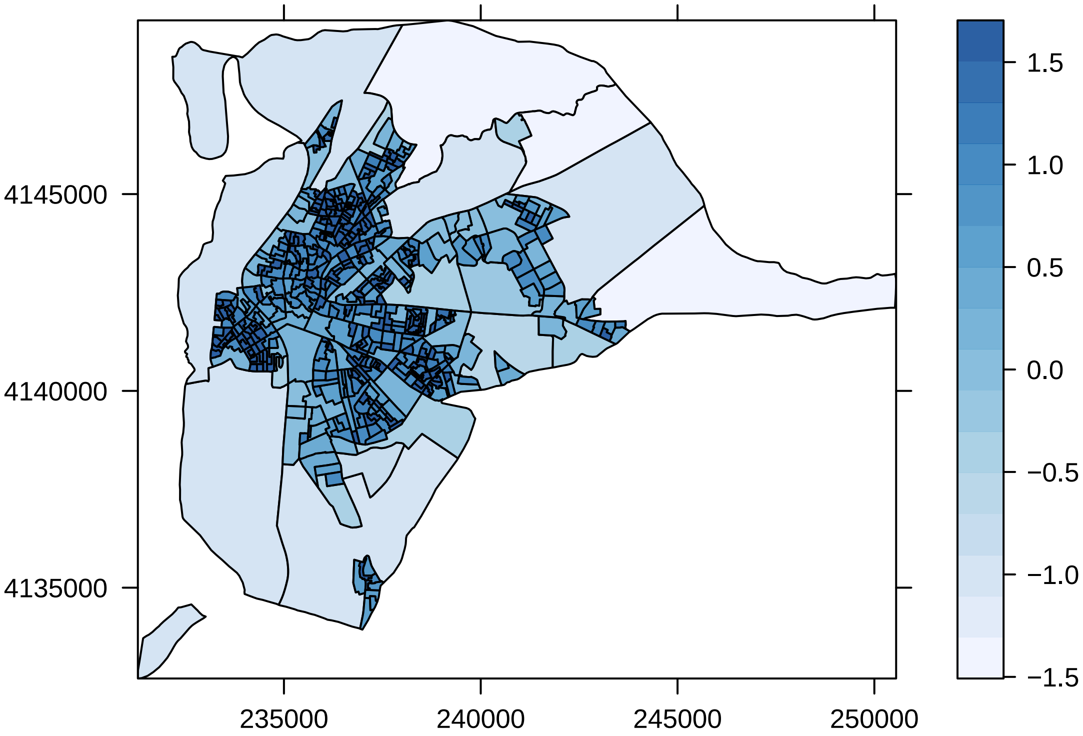
A second possibility, suggested by Fran, would have been to use population size by artificial surface which would account for the case that suburban areas in big census tracts might also be crowded.These tracts would become a higher weight, if we use the proportion of artificial surface as area base for the density. The problem is as indicated by the correlation plot is that there is almost no variation between the areas.

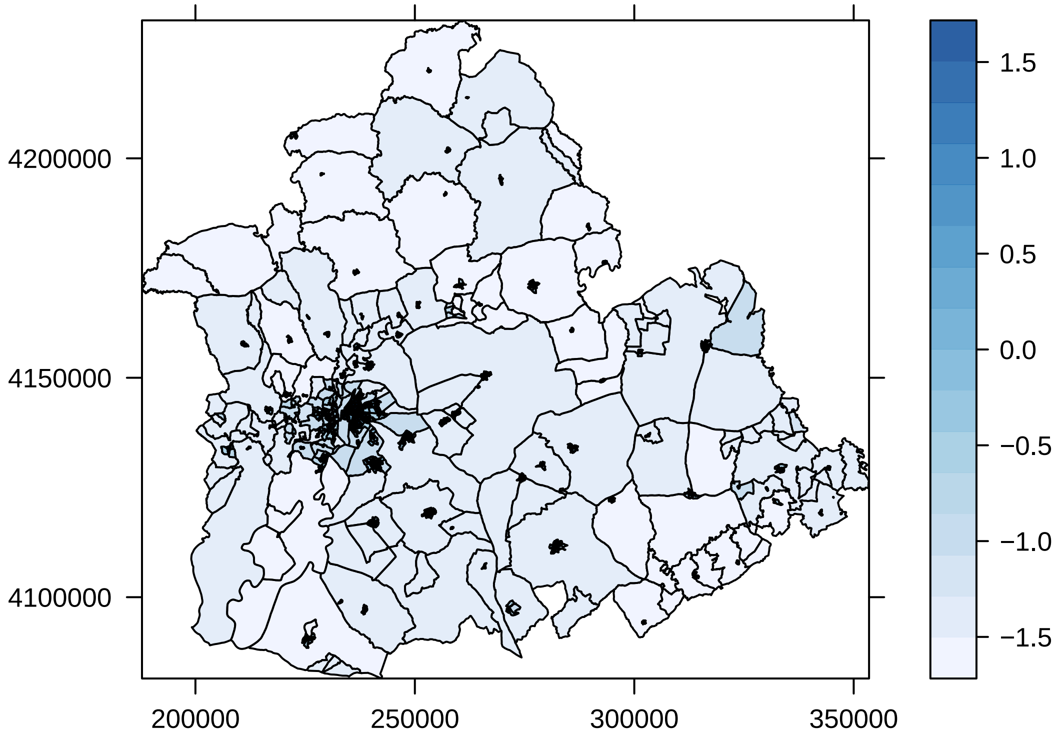


The two other physical variables, artificial surface proportion and road density, are probably the least questionable components when it comes to justification but at least the artificial surface has a peculiar distribution over the census sections with two peaks at the extreme ends.
The proportion of the population accounted service areas is not my favorite variable because we miss values for about 30% of the census tracts.
However, graphical tests indicate that the final indicator does seem to measure what we want it to do. It seemingly captures different degrees of “urban” especially when we look at the big cities.

```
## OGR data source with driver: ESRI Shapefile
## Source: "C:\Users\y4956294S\Documents\LONGPOP\Subproject 3 - Urban Environment and Health\RCode\Urba
## with 5381 features
## It has 8 fields
```

```
##
##   10
## 5381
```


4145000
4140000
4135000
235000
240000
245000
250000
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5



Please note, that this document is just for internal use. I am aware that we have to use GIS for mapping our indicators or other variables with geographical variation.

Environmental and social area features

Compared to the PAA paper, this is where we will have the biggest changes. I slightly disagree with Fran that the "Harmful Environment Indicator" is a good measure. In my opinion, the indicator we presented is a rather arbitrary measure which mixes different indirect effects which are hard to justify and combine in a meaningful way. Instead of the indicator, I am suggesting to include the additional area features in two model steps, one for physical environmental factors unrelated to urbanicity and one for population-based or social-environmental variables which all cover a slightly different aspect (or potential source of risk). For now, the physical environmental variables are cleanness, Pollution, and Noise. The social-environment is represented through number of delinquencies, percentage of employed, and percentage of single households (which might be related to urban environments). The choice was based on a few sensibility tests and is not exhaustive. Suggestions are welcome!

Spatial Autocorrelation

This is in fact something we have to reconsider as there are, to my knowledge no models on the market, with which we can easily estimate risks for individuals which are clustered in areas which features are spatially auto-correlated. Our model includes a random effect on the baseline for every census tract, but it does not account for proximity as far as I understood. Two solutions come to my mind (not exhaustive). 1.We apply the local Morons I only as measure to justify the multi-level structure - in way: Look! There is spatial variation so it does make sense to analyse mortality area differences at the census tract level 2.Or we can

go with Bivand et al (2013 - Applied Spatial Analysis with R) who state that “spatial patterning – spatial autocorrelation – may be treated as useful information about unobserved influences, but it does challenge the application of methods of statistical inference that assume the mutual independence of observations.”
I read somewhere that it is possible to use Lees L as justification for not accounting for the spatial autocorrelation in the regression model. This would include a little bit more research (see Lee)

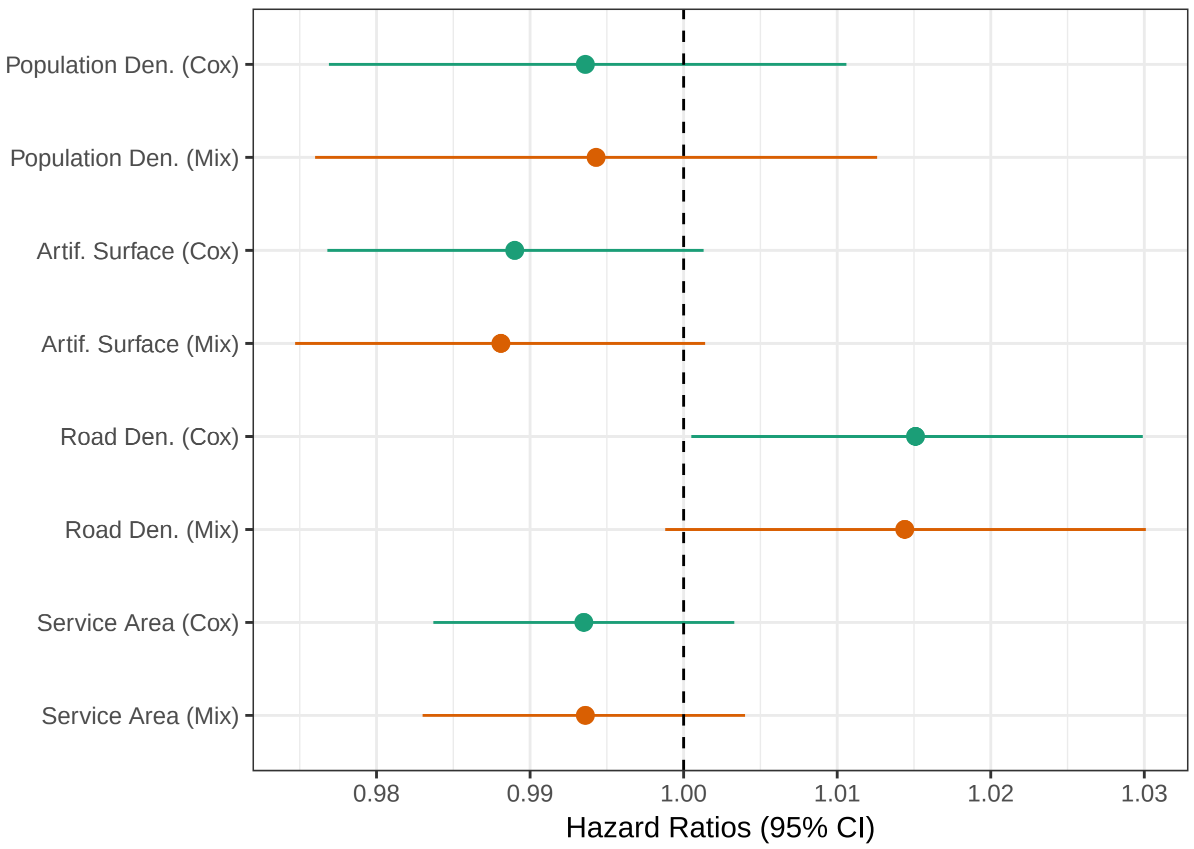
Model Results

As our focus is on area effects, I agree with Fran that the “urbanicity” indicator has to be in every model (That was already included in the paper I finally uploaded to the PAA website). I re-ran the analysis with the adjustments we made (new age groups, missing values etc.) and attached the results below. The table is unfortunately only visible in the pdf-document (excuses for the inconvenience. That is something I need to figure out later).
In contrast to the results presented at PAA, we include the additional area effects in single (hopefully) meaningful model steps. All models seem to fit the data better than the less complete model and the same model without random effects. This was tested with Log-Likelihood Ratio Test Statistics (see Thernau 2015 “Mixed Effects Cox Models”). In model 2 the environmental effects are incorporated and as we see in all models, the only consistent effect seems to come from contamination. The third model step accounts for additional population-based or social environment effects which are suggested to be all highly significant even when individual differences are included. The effect of delinquency seems not surprising, but also employment and household composition effect hold. And model 4 is the full model which is the most interesting since the effect of urbanicity disappears while other area features, especially the social-environment ones remain to have an impact.
Note: I will follow Frans advice to include interaction effects.

One of the comments after the PAA presentation indicated that it would be interesting for the reader to get to know about the effects of the single indicator components. We could include a second regression table with how the indicator components affect mortality in the single models or a forest plot as this preliminary one below:

Table 1: Cox PH Model with mixed effects

	Dependent variable:			
	Hazard Ratios			
	(1)	(2)	(3)	(4)
"Urbanicity" Index	1.0060 (0.996, 1.016)	0.9978 (0.986, 1.010)	0.9736*** (0.9606, 0.9866)	0.9938 (0.9812, 1.0064)
Perceived Cleanness		1.0008** (1.0002, 1.0014)	1.0001 (0.9995, 1.0007)	1.0000 (0.9994, 1.0006)
Perceived Pollution		1.0024*** (1.0014, 1.0033)	1.0020*** (1.0011, 1.0029)	1.0016*** (1.0008, 1.0025)
Perceived Noise		0.9992. (0.9983, 1.0001)	0.9986*** (0.9977, 0.9995)***	0.9992. (0.9984, 1.0001)
Delinquency			1.0019*** (1.0013, 1.0026)	1.0017*** (1.0011, 1.0023)
% Employed			0.9949*** (0.9937, 0.9960)	0.9979*** (0.9967, 0.9990)
% Single HH			1.0062*** (1.0036, 1.0087)	1.0070*** (1.0045, 1.0095)
Male				2.0900*** (2.0709, 2.1091)
Reference: Female				
Physically Dependent				3.0173*** (2.9506, 3.0839)
Reference: No Dependency				
Single				1.4052*** (1.3754, 1.4350)
Widowed				1.1809*** (1.1560, 1.2059)
Divorced/Separated				1.4708*** (1.4162, 1.5254)
Reference: Married				
No or Incomplete Educ.				1.3890*** (1.3477, 1.4304)
Primary/Secondary Educ.				1.1622*** (1.1181, 1.2063)
Reference: Tertiary Educ.				
Does not Own House				1.1379*** (1.1114, 1.1644)
Reference: Does Own House/Apartment				
Does not Own a Car				1.2623*** (1.2425, 1.2820)
Reference: Does Own Car(s)				
Observations	348,694	348,694	348,694	348,694
Log Likelihood	−541,809.1	−541,781.7	−541,701.9	−537,517.3
LR Test	111.8 (df = 2)***	54.8 (df = 5)***	159.6 (df = 8)	8369.2*** (df = 17)


Population Den. (Cox)
Population Den. (Mix)
Artif. Surface (Cox)
Artif. Surface (Mix)
Road Den. (Cox)
Road Den. (Mix)
Service Area (Cox)
Service Area (Mix)
0.98
0.99
1.00
1.01
1.02
1.03
Hazard Ratios (95% CI)