National Centre for Research Methods Working Paper

07/13

Simple, multiple and multiway correspondence analysis applied to spatial census-based population microsimulation studies using R

Didier Leibovici and Mark Birkin, TALISMAN node, University of Leeds

NCRM
National Centre for
Research Methods

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

Simple, multiple and multiway correspondence analysis applied to spatial census-based population microsimulation studies using R

Didier G. Leibovici[1] Mark Birkin[2]
[1] Nottingham Geospatial Institute
University of Nottingham
[2] Centre Spatial Analysis and Policy
School of Geography
University of Leeds

Abstract:
As a bivariate and multivariate multidimensional exploratory method, simple and multiple correspondence analyses have been used successfully in social science for survey or questionnaire results descriptions. Nonetheless, the complexity of social interactions including health status indicators, with also the need to take into account the spatial and temporal realm of the survey, may incline to look at variable associations in a multiway approach instead of a two-way matrix analysis. This means for example, that interaction of order three between the spatial configuration (say the Output Areas of an urban zone), the set of categorical variables (say selected from a census survey) and the evolution (say every 5 years over a 30 years period) would be considered in order to differentiate spatio-temporal associations across categorical variables. For census-based spatial simulation models such as microsimulations, exhibiting this kind of properties is useful as forecasts moves of population characteristics to be considered for healthcare policy scenario analysis. In this paper it is shown how to run this type of analysis within R using a package dedicated to multiway analysis (the R package PTAk), that is, working on multi-entry array data using an algorithm extending classical multidimensional analysis. A didactic approach from two-way analyses to multiway ones, of the same dataset generated from a population spatial simulation model allows a critical demonstration of the potential of the different t methods. Particular attention is also given to the different choices of spatial units and the scale variation effect within a nested administrative zoning system that can be analysed by a correspondence analysis with respect to a model (extending the approach using the independence model) and which can be done for a simple, multiple of multiway correspondence analysis.

Keywords: microsimulation, population simulation, census data, health indicators, spatial units, correspondence analysis, multiway analysis, data reduction methods, multi-entries arrays, tensor, R

1 Introduction

Factorial Correspondence Analysis (FCA) allows breaking down, in a multidimensional analysis way, the residual to the probabilistic independence for the joint distribution of two categorical variables. Statistical properties of this method, linked to some matrix algebra properties, leads naturally to extend it to more than two qualitative variables with the Multiple Correspondence Analysis (MCA) decomposing a statistic expressing all pairwise residuals from independence (see Lebart et al. (1984) or Murtagh (2007) for a more matrix algebra exposé, Le Roux and Rouanet 2010 for historical point of view and a geometrical focused approach). Therefore, associations of variable categories measured from a population survey, highlighted from the matrix decomposition as contributing to large residuals, are depicted, described and interpreted within a socio-demographic analysis of the "clouds" of points (individuals and categories) (Le Roux and Rouanet, 2008). In social sciences and other disciplines dealing with contingency tables analysis such as coming from surveys and questionnaires, MCA has been very successful as a tool to explore potential variable associations and generate hypothesis as well as capturing and summarising dimensions (also called scores, indices, factor scales) of a social, demographic or behavioural domain and in other population survey sciences (*e.g.,* Savage, 2010, Le Roux and Rouanet 2008, Gatrell et al. 2004).

Complex interaction can be revealed using this method but only implicitly from a list and pairwise interactions, and after a purely theoretical interpretation from external sources related to the studied domain. In other words if a three-way interaction of categorical variables is to be analysed, the method to perform this has to be based on analysing the joint distribution of the three variables involved: a three-way entries contingency table. This is further called a multiway approach and each entry of the multiway table is called a mode. Multiway approaches are obvious approaches when a repetition is part of the design of the study. For example a spatio-temporal survey study (longitudinal study with a geographical anchor) in which answers to a questionnaire at regular time interval are collected over a spatially sampled population over a certain period, can be summarised in a table of counts with three entries: the spatial mode made of one variable identifying the spatial unit, the item mode identifying the categories of each variable and the time mode identifying the date or sub-period. More entries can be considered if cross-tables with one or more categorical variables are chosen, so that a mode can be semantically describing a particular sub-domain, sub-scale, aspect present in the list of variables measured. If multiway approaches can be useful to exhibit a complex interaction, higher than second order, the limitations are dictated by sample size (to get potentially enough observations for each combination of categories of crossed variables) and presence of structural zeros (*i.e.,* either impossible or very unlikely combinations of categories of different variables, *e.g.*, an occupation typically found in young females will generate a bunch of zeros if the variable is crossed with age and then crossed with sex (to get a *occupation x age x sex* table).

Correspondence analysis can be used for any type of studies with categorical variables, the focus in this paper will be given to spatial microsimulation population studies derived from census data. In order to be used for policy scenario analysis, the spatio-temporal variation of associations is the main interest in these types of studies as well as the potential interference of the choice of spatial units. The natural

spatial evolution of the population, or its aspects under a “what if” scenario, are urban forecasts needed as the basis for policy decision-making (Birkin and Clarke 2011).
The MoSeS dynamic MSM is providing data for the analysis and illustration of potential methods to further analyse MSM results in general; the features of this simulation are described next. Starting from a situation in which demographic attributes are assumed to vary simultaneously, it will be shown how this null model can be refined using the methodology of correspondence analysis. Many statistical software have inbuilt functionalities to perform correspondence analysis or very similar methods multidimensional methods (*e.g.,* SPSS, SAS, MiniTab, **R**): **ca** (Nenadic and Greenacre, 2007), **ade4** (Dray and Dufour, 2007) to name a few packages within the **R** software (**R** Development Core Team 2009).

The paper aims at illustrating the multiway approach in progressing from simple and multiple correspondence analysis whilst using practical hands on **R** coding to use. Factorial Correspondence Analysis on k-modes tables (FCAk) (Leibovici and El Maâche 1997,) has been proposed as a potential multiway extension of the FCA when considering k>2 categorical variables. The R package **PTAk** enables the FCAk to be computed among other multiway methods (last version of the package is described in Leibovici 2010).

2 Spatial microsimulation and MoSes

The technique of microsimulation (MSM) is well-established as an approach to econometric evaluation and policy analysis (Orcutt, 1957). The essence of MSM is the representation of populations (e.g. of customers or economic agents) as lists of individuals rather than an array of the counts of different group members. In this way the full variety of the population can be captured effectively (Birkin and M. Clarke, 2011), which is particularly important in geographical contexts in which spatial differentiation is an important consideration; and particularly so when some element of spatial interaction between areas renders demographic arrays to be explosively large (see for example, van Imhoff and Post, 1998). Hence spatial microsimulation has become a popular mechanism for the analysis of all kinds of socio-demographic problems in a geographical context, ranging from healthcare (Smith et al, 2006, Procter et al, 2008) to labour markets (Ballas and Clarke, 2000), poverty (Tanton et al, 2009) to retail (Birkin and Clarke, 1987; Nakaya et al, 2007).

Whilst MSM are of interest as analytical and synthetic estimation tools, a particularly important sub-class of approaches are dynamic MSM, in which individual representations are used as a powerful basis for demographic projection. Methods include comparative static approaches in which updated populations are synthesised from marginal counts (Ballas et al, 2005) and models which estimate the time between changes in demographic states e.g. associated with fertility, illness, mortality or migration (‘event-based models’, e.g. DynaCan – see Statistics Canada, 2009) as well as transition-oriented approaches such as the MoSeS model which is

discussed further below, and in which the probability of evolution between states is evaluated across discrete time periods.
An important feature of dynamic microsimulation models is their capacity to generate extremely large and rich datasets. Consider for example a demographic model of the city of Leeds in which the population is segmented into 2,439 geographical neighbourhoods, 30 time periods and 10 characteristics, each with six attribute categories (see below). The number of potential elements in this simulation is the order of 4 million (i.e. 2,439 x 30 x 10 x 6). There is an analytical requirement to identify patterns in this data set for a number of reasons. First, one might wish to locate key trends in space and time – for example, where is the impact of demographic ageing most significant, and when do these trends begin to accelerate dramatically? Secondly, forensic investigation can support tests and demonstrations of the robustness of the underlying models – where are the anomalies, outliers and unusual cases, and how can these be explained and rationalised? Thirdly, when the purpose of simulation is to inform policy appraisal or 'what if?' experiments (as will often be the case) then it may be useful to explore whether different interventions lead to pronounced or significant changes in the distribution of an outcome variable.

For problems in which data varies in a single dimension then techniques for the identification of patterns or discrete changes, are well-established. In many cases something as straightforward as simple regression could be sufficient. Similarly in two dimensions then methods such as spatial clustering might provide a robust approach to pattern recognition. However in the current context in which the system of interest varies simultaneously in at least three dimensions (i.e. a characteristic is differentiated by attribute variation in both space and time) then the problem is more challenging.

The MoSes[1] project (*Modeling and Simulation for e-social science*, Birkin et al. 2009) promotes demographic simulation using hybrid models. Whilst grounded in the methods of microsimulation, concepts from spatial interaction modelling and agent-based systems are incorporated into the MoSes approach (Townend et al. 2009, Wu and Birkin 2012). Geographical microsimulation involves the fusion (matching, merging) of census and survey data to simulate a population of individuals within households (for different geographical units), whose characteristics are as close to the real population as it is possible to estimate (Ballas et al. 2005). To generate the baseline spatial data different techniques exists for optimization of the fusion and resampling using usually re-weighting methods (see a recent review in Hermes and Poulsen 2012). For population simulation forecasts, MoSes approach uses dynamic modelling for population projections where processes such as marriage, fertility, or mortality are operated individually rather than resampling the population for the new period as when using static methods *(i.e.,* derived from macro, aggregated parameters) (Birkin and Clarke 2011).

[1] MoSes is now part of the Generative e-social spatial simulation GENeSIS initiative http://www.genesis.ucl.ac.uk/

Initial composition of households microdata have been generated using MoSes based on the Sample of Anonymised Records (SARs) from the 2001 UK population census and resampled according to local distributions (in each Leeds Output Areas) provided by the SAS (Small Area Statistics) issued from the census data. The final dataset covers the Leeds area from 2001 with 296769 households composed of 701219 individuals to 2031 with 351953 households representing a population size of 826015. Each individuals has attributes related to the economical, demographic, and health domains as described in the survey data (SARs), *e.g.:*

*(**age**)*	*Age of the respondent in 5 intervals up to **age18, age26, age44, age64** and **age80***
*(**sex**)*	*Male or female labeled **sex1** and **sex2***
*(hr**soc**grd)*	*Social grade of household reference person labeled **soc-9-soc5** (see Appendix)*
*(**car**)*	*Number of cars owned labeled **car<value>,** with values **0, 1, 2**(more than 1)*
*(**eth**ew)*	*Ethnic group for England and Wales reclassified in 5 groups (see Appendix)*
*(**hea**lth)*	*Self-assessment for general health in the last 12 months: -9- not applicable (student living away), 1-good, 2-fairly good, 3- not good (here labeled **hea-9 – hea3**)*
*(l**lti**)*	*Limited long term illness: 1-yes, 2-no (here labeled **llt1** and **llt2**)*
*(h**hlt**hin**d**)*	*Household health and disability indicator: 0-noone with general health as 'not good' or llti, 1- at least one has either (here labeled **hlt-9, hlt0, hlt1**)*
*(**pro**vcar**e**)*	*Number of hours of care provided per week **pro<value>** (see Appendix)*
*(**tr**anwrk0)*	*Transport to work (labeled **trw<value>)** regrouped in 4 classes (see Appendix)*

Not all the variables have been used here but the covered domains mentioned above are representing altogether 96 categories (they will be described as they will be highlighted during the analyses, see further sections, see also the annexe). The analyses will be made on aggregated data at different scales: ward level (33 wards), middle layer Super Output Areas MSOA (108 MSOA), lower super output areas (476 LSOA) and at Output Areas levels (2439 OAs). Seven time steps have been retained for the analysis: every 5 years from 2001 to 2031.

3 Simple and Multiple Correspondence Analysis

The basics of correspondence analysis and its extension to multiple correspondence analysis help to understand the multiway extension; further explanations can be found in the references quoted in introduction. With two categorical variables respectively with I and J categories, a simple contingency table or cross-table n_{ij} with $i = 1\ to\ I$ and $j = 1\ to\ J$ leads to the joint distribution of the two qualitative variables: $p_{ij} = n_{ij}/N$where $N = \sum_{ij} n_{ij} = n_{..}$ is the total sample size. Testing the independence of the two variables, that is comparing p_{ij} with $\hat{p}_{ij}$=$p_{i.}p_{.j}$, can be performed using the chi-square statistic, measuring the lack of independence by:

$$\chi^2/N = \sum_{ij} p_{i.}p_{.j}(\frac{p_{ij}-\hat{p}_{ij}}{p_{i.}p_{.j}})^2 \qquad (1)$$

Equation (1) is also the inertia, that is, the sum of the eigenvalues, of a particular principal component analysis (PCA) decomposition, in fact a singular value decomposition (SVD), of the $I \times J$ matrix $Z, Z_{ij} = \frac{p_{ij}-\hat{p}_{ij}}{p_{i.}p_{.j}}$ with diagonal metrics D_I and $D_J$made from the margins respectively $p_{i.}$'s and $p_{.j}$'s of the matrix *P* (for more details see for example Greenacre 2007) This multidimensional analysis, the FCA, can therefore provide a break down of the lack of independence between the two variables with quantitative evaluation of the revealed associations (the eigenvalues and percentage of inertia they represent) and with scores for each categories on each dimensions (eigenvectors) linked to their contribution to the departure from

independence. Because the $I \times J$ matrix $P = {}^t(V_I)V_J$, where V_I is the $N \times I$ matrix identifying the categories for each of the *N* observations (individuals) and *idem* for V_J, the FCA of *P* and the FCA of the concatenation these two matrices of 0's and 1's, that is the FCA of $U = (V_I \vdots V_J)$ or also the FCA of $B = {}^t(U)U$ (called the Burt table) have the same results for the scores (re-standardised) of each categories and the eigenvalues have a simple relationship: $\lambda_U = \sqrt{\lambda_B} = (\sqrt{\lambda_P} + 1)/2$.
When *U* is a concatenation of more than two variables, the FCA of *U* is called the MCA as well as the FCA of B, but the latter is often just called the FCA of the Burt table.
Within the **R** software, and besides the `corresp()` function in the **MASS** library (by default in the **R** installation), packages such as **ca**, **ade4** and **PTAk**, allow to perform the FCA and MCA methods as well a few other packages: see the CRAN task view on Multivariate Statistics at your local CRAN[2]).

3.1 data manipulation in R

After reading the data files of the separated values, the table(s) are prepared and the correspondence analysis can be performed. A typical coding is given below and the R script file can be found at c3s2i.free.fr/MoSesOAWard.FCAk.tutorial.R.

3.2 a simple example

Reading the year 2001 data of the Leeds microsimulation from the MoSes project into a `data.frame` and creating a contingency table from two categorical variables to perform the correspondence analysis is done by:

```
> dat01 <- read.table("trans01.txt")
> crossTab. sochealth <- xtabs(~ V19 + V13, exclude = c(NA, NaN),data = dat01)
> library(PTAk)
> crossTab.sochealth.FCA2 <-FCA2(crossTab.sochealth)
summary(crossTab.sochealth.FCA2)
 +++ FCA-  complete independence  ++  2 modes+++
     ++ Contingency Table  crossTab.sochealth   6 4  ++
     -----Total Percent Rebuilt---- 100 %
     ++ Percent of lack of  complete independence  rebuilt  ++  100 %
                    selected pctoafc >  0.5 %  total=  100
    -no- --Sing Val-- --ssX-- --Global Pct--  --FCA--
vs1    1     1.000000  1.0697      93.485349        NA
vs2    2     0.259736  1.0697       6.306769 96.80900
vs3    3     0.041806  1.0697       0.163389  2.50802
vs4    4     0.021816  1.0697       0.044494  0.68298
++++              ++++
 Shown are selected  over  3  PT  with pct FCA > 0.5 %
```

Each line with a ">" is an instruction within the R interface which can return or not a printed result. The first three lines are executing something and storing the result in an object (on the left hand side of the "<- " symbol). V19 is the *hrsocgrd* categorical variable with 6 categories and V13 is the *health* variable. `FCA2()` is the implementation of an FCA within the R package **PTAk**, loaded by the command/function `library()`. Because of the linear relationships: $p_{.j} = \sum_i p_{ij}$ and *idem* for the row margins, decomposing the matrix Z or simply the matrix of the $\frac{p_{ij}}{p_{i.}p_{.j}}$ is equivalent. This is apparent from the first singular value, `vs1`, to be `1.000000` and

[2] Mirrors of Comprehensive R Archive Network (CRAN) are at www.r-project.org *e.g.,* http://cran.ma.imperial.ac.uk/ is one in the U.K.

associated components to be vectors of 1's naturally retrieved within the algorithm. The components are accessed from the output object, basically a list of results for each mode (entry of the table), so here `[[1]]` corresponds to the rows of the cross-table and `[[2]]` to the columns; `$v` is the matrix of components values for each singular value:

```
> crossTab.sohe.FCA2[[1]]$v[1,]
 [1] 1 1 1 1 1 1
> crossTab.sohe.FCA2[[2]]$v[1,]
[1] 1 1 1 1
```

This choice of implementation is useful for first quantifying directly from the analysis the amount of independence between the two variables (for this FCA2 analysis `93.5%` of the variability is explained by independence), and will be also useful for the extension to *k* variables, the FCAk method. Traditional plots in FCA or MCA involves biplots, in PTAk unscaled scatterplots (raw vectors are normed to 1 within their metric space) are the default:

```
> plot(crosstab.sohe.FCA2,nb1=2,nb2=3,mod=c(1,2))
```

A symmetric-map biplot (as the default plot in **ca**) as represented in Figure 1 can be performed by:

```
> plot(crossTab.sochealth.FCA2,nb1=2,nb2=3,mod=c(1,2),
       coefi=list(c(0.259736,0.259736),c(0.041806,0.041806)) )
```

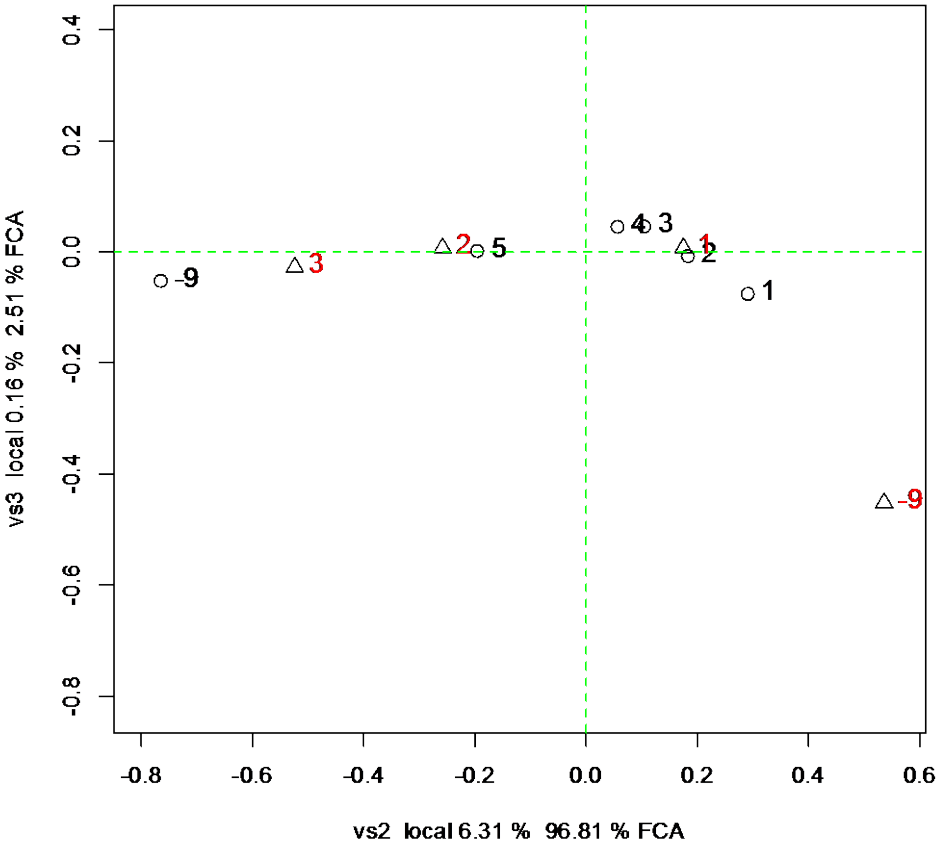

Figure 1: Symmetric-map biplot for the FCA2 of the *nssec x health* table.

where `coefi` is weighting each element of the components to display. Figure 1 expresses a gradient of health on the first dimension with a minor moderation expressed on the second dimension (only 2.5% of departure from independence), associated with a gradient on the social grade (not applicable social grade being more likely at the end of this gradient associated with poor health). This quick interpretation has to be “weighted” by the quality of projection of each category and their contributions on each dimension. This is done using the classical measures COS2 and CTR, which are respectively, the ratio to the whole variability of a given category (an item within a mode) and the ratio to the variability accounted by the component (the squared singular value):

```
> cbind(CTR(crossTab.sochealth.FCA2,mod=1,solnbs=2:4),
           COS2(crossTab.sochealth.FCA2,mod=1,solnbs=2:4, FCA=TRUE))
   ctr_:2:_vs2 ctr_:3:_vs3 ctr_:4:_vs4 cos2_:2:_vs2 cos2_:3:_vs3 cos2_:4:_vs4
-9         535          95         304          991            5            4
1          190         496           0          936           64            0
2           84           6          44          995            2            4
3           25         183          34          832          160            8
4            9         220          57          592          381           27
5          158           0         560          976            0           24
> cbind(CTR(crossTab.sochealth.FCA2,mod=2,solnbs=2:4),
      COS2(crossTab.sochealth.FCA2,mod=2,solnbs=2:4, FCA=TRUE))
   ctr_:2:_vs2 ctr_:3:_vs3 ctr_:4:_vs4 cos2_:2:_vs2 cos2_:3:_vs3 cos2_:4:_vs4
-9          34         927          32          582          414            4
1          294          20          38          997            2            1
2          234           7         522          984            1           15
3          439          46         408          991            3            7
```

From their definitions the sum of CTRs across the mode items is 1000 and the sum of COS2 across the whole components rebuilding all the data is also 1000. The social category gradient is weakened by the small CTRs for category 3 and 4 on vs2, which are nonetheless well represented on it. The second dimension expresses more the opposition of category 1 to 3 and 4, associated with the -9 for health (not applicable, student living away), which builds this dimension (CTR = 927).

3.3 an MCA on spatial data

Besides being able to analyse more than two variables, one advantage of the MCA approach is to obtain the subject “cloud of points” as well as the variable-categories scatterplots. Here there is no desire to derive loadings for 700 000 people as we rather use directly the SARs data for this, but after the simulation one may be interested to know the spatio-temporal variations over the Leeds area. Notice that according to the distributional equivalence property of the chi-square distance (see for example Lebart et al. 1984), performing an MCA where the subject entry has been aggregated to spatial units would be equivalent to the non-aggregated one if the profiles were proportional. This is the implicit assumption when aggregating data.

The variables spatially analysed chosen were *hrsocgrd* , *sex, age, ethew, health, hhlthind, llti, provcare, tranwrk0* and *carh* so covering socio-demographic information including ethnicity, general health related information, and some transport use information (see appendix for a description of the categories). Below is the R code with the results to perform an MCA for these variables aggregated at OA levels (2439 OAs in Leeds), for the year 2001 (dat01). The OA variable is V3:

```
> spaTab <- xtabs(~ V3 + V19, exclude = c(NA, NaN),data = dat01)
> colnames(spaTab)=paste("soc",colnames(OA19),sep="")
> listV=c(30, 6, 7, 11,13,14,23,29,33)
> listVn=c("sex","age","car","eth","hea","hlt","llt","prov","trw")
> for (v in 1:length(listV)){
 temp <- xtabs(~ V3 + get(paste("V",listV[v],sep="")), exclude = c(NA, NaN),data =
 dat01)
colnames(temp)=paste(listVn[v],colnames(temp),sep="")
spaTab <- cbind(spaTab,temp)
}
> dim(spaTab)
[1] 2439   39
```

```
> spaTab.MCA <- FCA2(spaTab)
> summary(spaTab.MCA,testvar=2)
+++ FCA-  complete independence  ++  2 modes+++
     ++ Contingency Table  spaTab   2439 39  ++
     -----Total Percent Rebuilt---- 100 %
     ++ Percent of lack of  complete independence  rebuilt  ++  100 %
                    selected pctoafc >  2 %  total=  89.58517
    -no- --Sing Val-- --ssX-- --Global Pct-- --FCA--
vs1    1     1.000000  1.1038       90.59879      NA
vs2    2     0.224699  1.1038        4.57431 48.6566
vs3    3     0.118833  1.1038        1.27936 13.6085
vs4    4     0.097135  1.1038        0.85482  9.0927
vs5    5     0.077182  1.1038        0.53970  5.7408
vs6    6     0.067015  1.1038        0.40689  4.3280
vs7    7     0.058105  1.1038        0.30588  3.2536
vs8    8     0.053006  1.1038        0.25455  2.7076
vs9    9     0.047751  1.1038        0.20658  2.1973
++++                    ++++
 Shown are selected over 38 PT with pct FCA > 2 %
```

The independence (2 by 2 independence between the variables as observed on the OAs) captures 90% of the variability. This is represented by the margins of the table as displayed in Figure 2 illustrating respectively variation of the population in Leeds at OA levels and the observed proportions for each variable category over the whole area. The categories of age, ethnicity and transport to work have been recoded to ease the analysis (see appendix table).

To display the spatial component the R code is as follow, where `plotmapbarq()` uses the plot function from the package **sp** and builds up a legend based on percentiles of the values:

```
> library(RColorBrewer)
> Yl <- brewer.pal(7, "PuOr")
> library(maptools)
> OAmap = readShapePoly("Zleeds_oas.shp", IDvar="ons_label")
> met <- FCAmet(spaTab)
> plotmapbarq(Poly=OAmap,nclass=10,nvec=met$met[[1]]*(1/2439)*100,
            colrmp=colorRampPalette(Yl)(10))
```

Variables histograms are done using the following code:

```
> marginVar701219dat01=apply(spaTab,2,sum)    #
all(round(OAmet$met[[2]]*10*701219)==marginVar701219dat01)
> par(mfrow=c(3,1))
> barplot(marginVar701219dat01[1:6]/7012.19,xlab="hrsocgrd",ylab="%")
> barplot(marginVar701219dat01[7:8]/7012.19,xlab="sex",ylab="%")
> barplot(marginVar701219dat01[9:13]/7012.19,xlab="ageh",ylab="%")
```

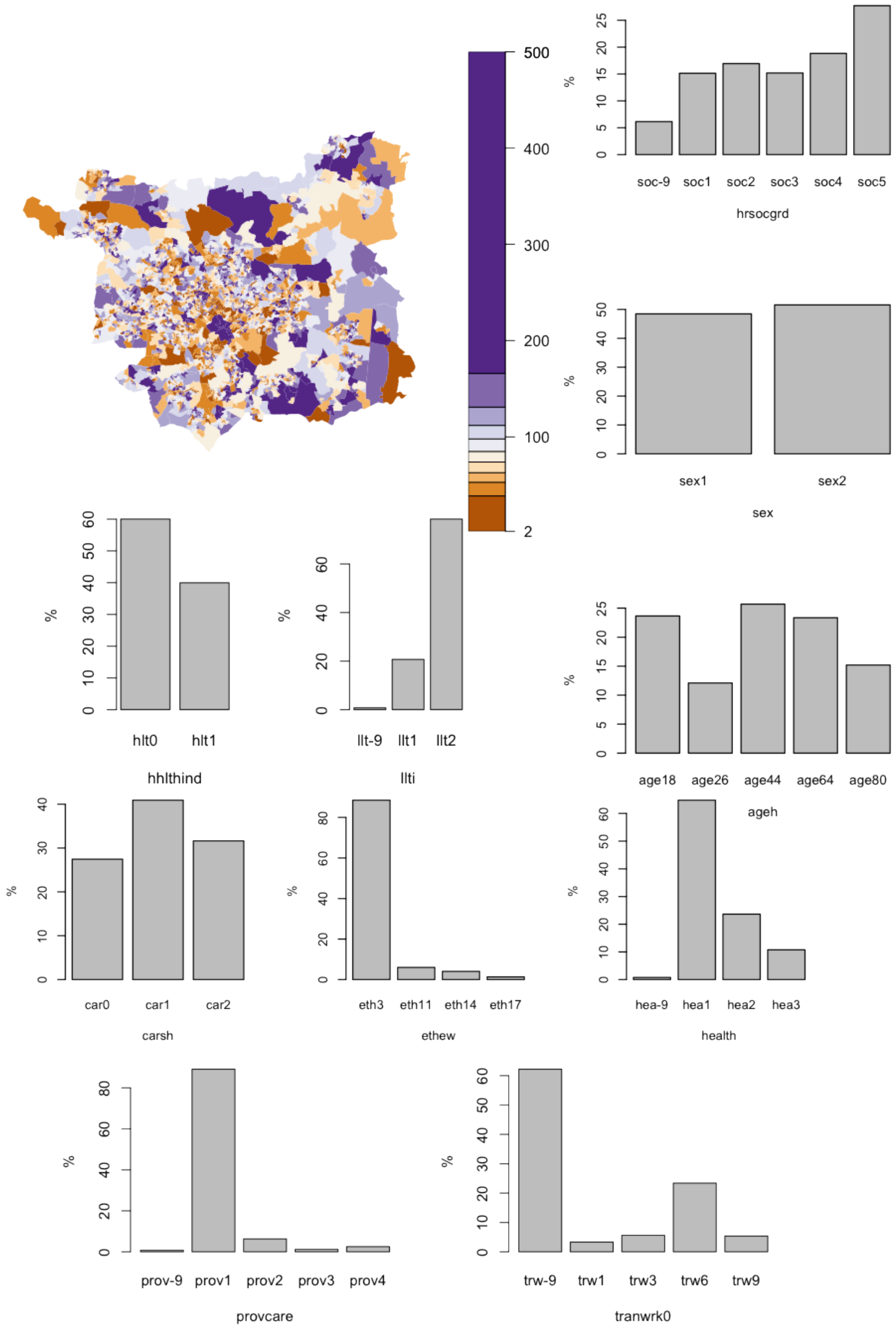

Figure 2: Margins –spatial margin represented at OA level in % relative to uniform distribution (100%≍1/2439 ≍290) where a threshold >500 was applied to be able to depict the distribution of %, the colours classes are spread along the 10 percentiles – variable margin split by variable representing for each their distribution in 2001 overall at Leeds.

Notice that in theory in Figure 2, the population distribution at OA levels should be relatively uniform, so around the 100% level. In theory UK census boundaries are made to count around 300 people in each OA, but in fact at Leeds the distribution from the 2001 census shows 75% of the OAs within the 255-325 range and also a minimum at 108 and a maximum at 1236. There is nonetheless a discrepancy between the 2001 census OA counts and the values from the MoSes microsimulation which is explained by a two stage strategy in the simulation: a highly constrained simulation at ward level (around 20 000 people) then a redistribution at OA level with less constraints (see Birkin et al 200x for more details).

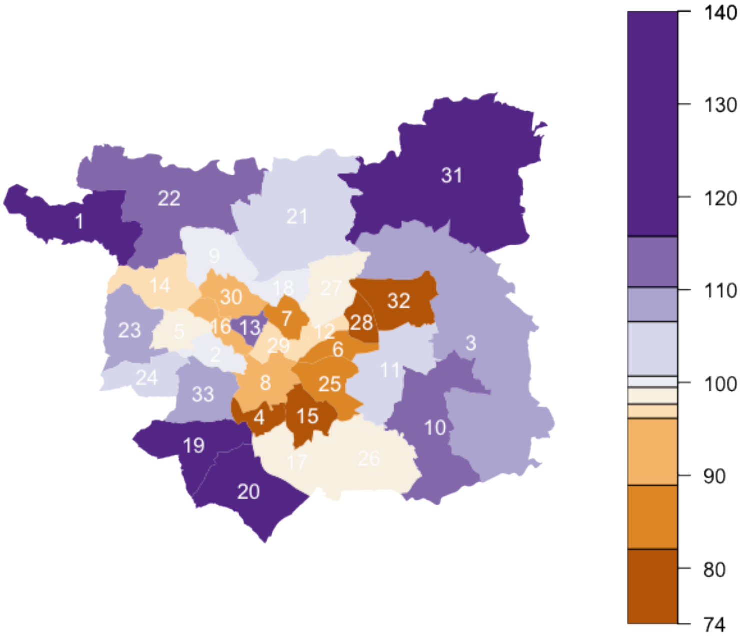

Figure 3: Margins –spatial margin represented at ward level in % relative to uniform distribution (100% ≍ 1/33 ≍ 21500) – ranging from 16 000 to 29 000 besides Headingly (University area- 17) the wards in the centre are less populated and wards at the southwest, northwest and northeast are the most populated. (1-Aireborough, 2-Armley, 3-Barwick and Kippax, 4-Beeston, 5-Bramley, 6-Burmantofts, 7-Chapel Allerton **8-City and Holbeck**, 9-Cookridge, 10-Garforth & Swillington, 11-Halton, 12-Harehills, 13-Headingley, 14-Horsforth, 15-Hunslet, 16-Kirkstall, 17-Middleton, 18-Moortown, 19-Morley North, 20-Morley South, 21-North, 22-Otley & Wharfedale, 23-Pudsey North, 24-Pudsey South, 25-Richmond Hill, 26-Rothwell, 27-Roundhay, 28-Seacroft, 29-University, 30-Weetwood , 31-Wetherby, 32-Whinmoor, 33-Wortley)

Looking at the CTR and COS2 together with Figure 4, the vs2 dimension opposes social grade (*hrsocgrd*) category soc5 (E- on benefit/unemployed) in association with young adults (age26- over 18 and less/equal to 26 years of age) from ethnicity groups 11 14 or 17 (coded here as Asian, black Caribbean, and other in a lesser extent) with no car (car0), to, the social grades soc1-3, old adults (age64- over 44 and less/equal to 64), and being from white origin (eth3 ethnicity groups 3) with more than one car (car2). The transport to work categories trw1, trw6 and trw9 (respectively public transport, personal/taxi motoring, cycling or on foot) are also well projected on vs2 to this latter side but trw6 is mainly illustrating (CTR and COS2) the opposition with the not applicable category trw-9. On the vs3 dimension, the social grades soc1-4 are spreading like a gradient, opposing soc1-2 to soc3-4 along with the not applicable category soc-9.

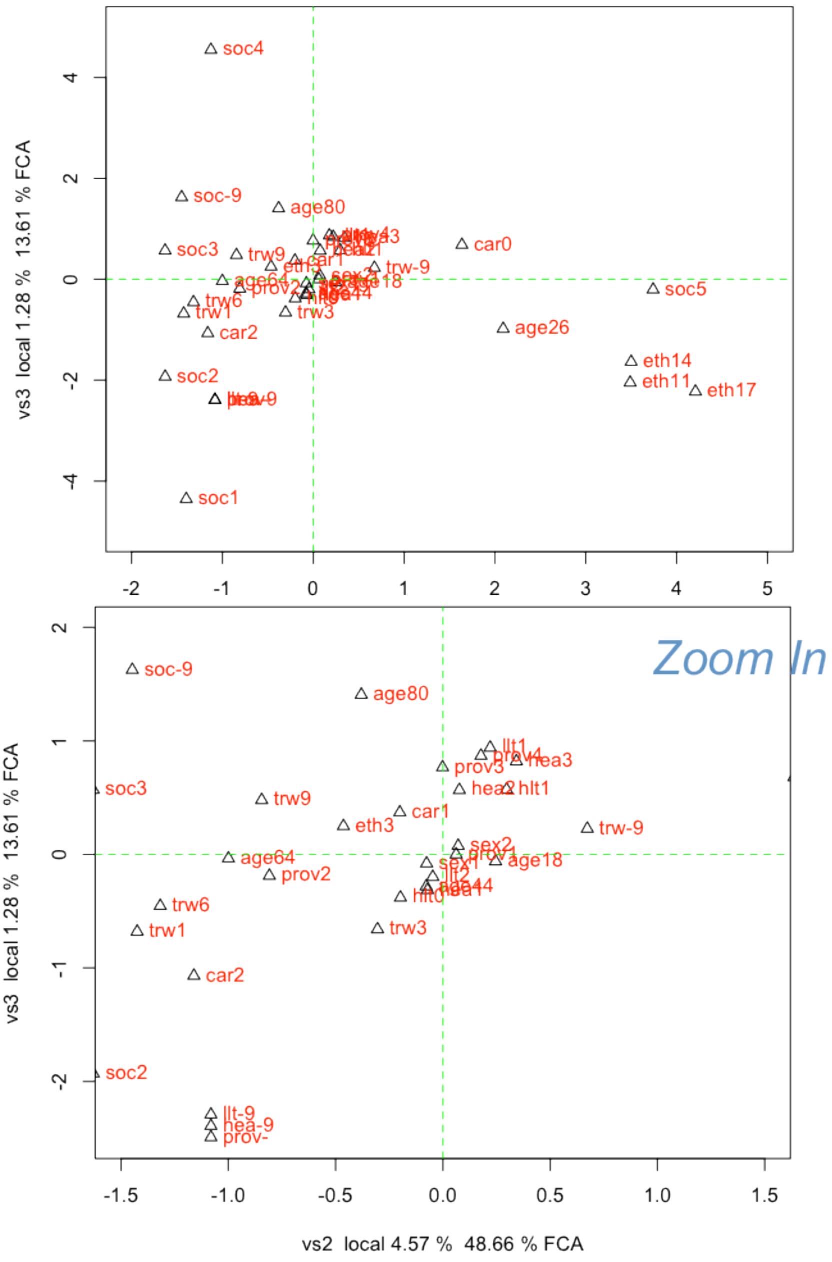

Figure 4: MCA of the OA aggregated chosen set of variables - first two variables components representing altogether 62% of the lack of independence (the bottom picture is a zoom in of the top one)

The CTRs are nonetheless much higher for soc1 and soc4; soc3 is not well projected on vs3 (low COS2). This is associated with more than 1 car (car2) in the soc1-2 side and with good health (hea1, hlt0, llt2; note that the Not Applicable or -9 category for llti, health and provcare are also on this side) as opposed to soc4 (category D: semi-skilled and unskilled manual workers) with age80 (over 64 years of age) with limited long term illness, fairly good or not good health and needing a fairly high number of hours for care (llt1, heal2-3, prov3-4).

Spatially (see Figure 5) the above descriptions for the vs2 dimension are more likely associated: on the positive side with the Northwest close to the city centre (wards 13, 29 and including the Northeast of 8 (city centre)) and on the negative side of vs2 with a few Northern and a few Eastern rural OAs (East of ward 22, West of ward 21, West of ward 31, West of ward 3, a good part of ward 10). For the vs3 dimension, the positive side is expressed on the East and South of the city centre (covering parts of ward 25, 15, 17 and Eastern part of 20), and the negative side is also on the Northern part of the district and more in the wards 27 and Western parts of 3 and 31.

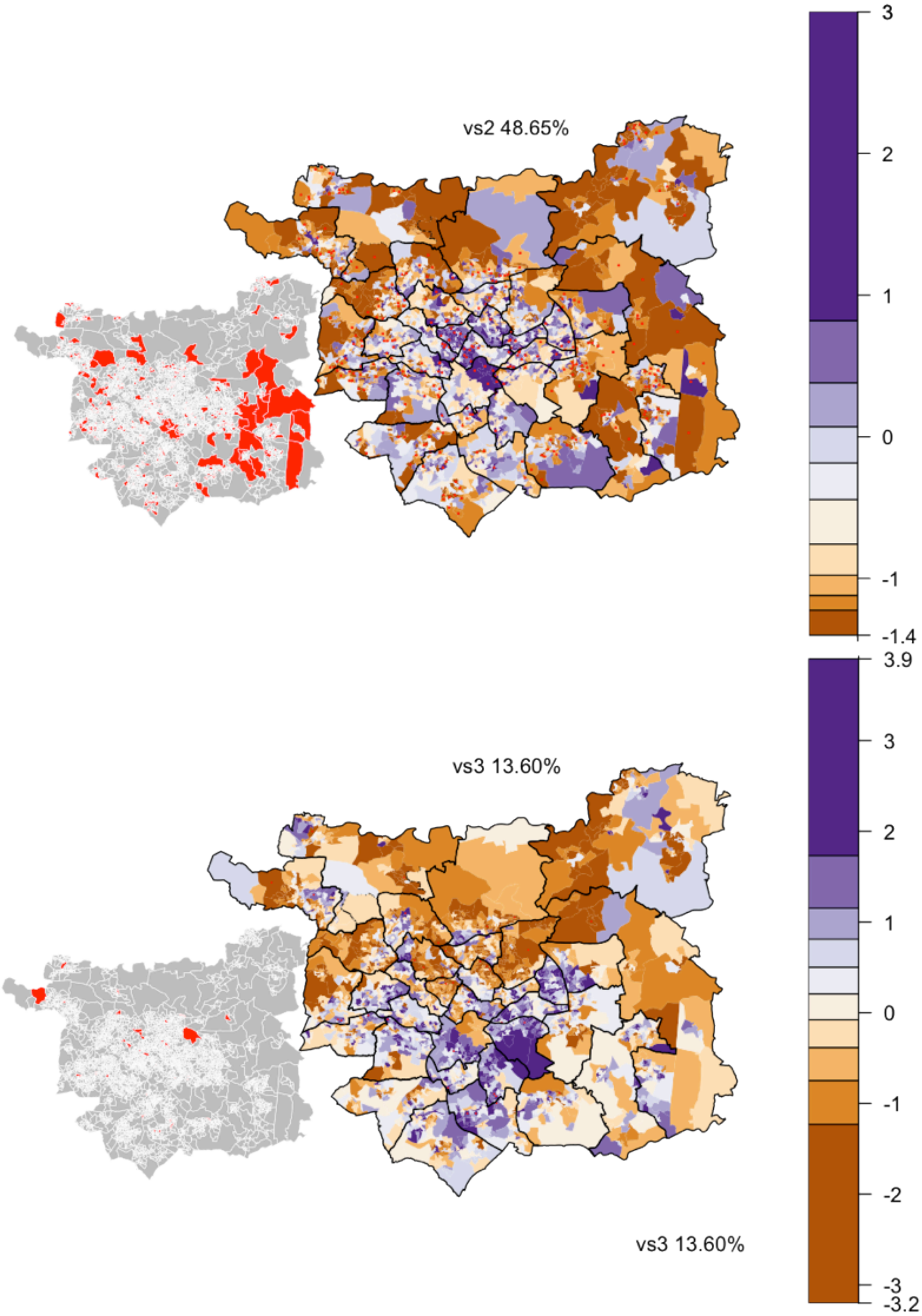

Figure 5: MCA of the chosen set of variables aggregated at OA level - first two spatial components representing 62% of the lack of independence. OAs with the highest CTR and COS2 have been tagged with a red "." and in the grey thumbnail map (on the vs2 map 306 OAs and on the vs3 map 34 OAs)

The overlay and tagging on the plots are done using the following R code, here for the spatial component vs2:

```
> poilab=rep("",2439)
> poilab[CTR(spaTab.MCA,mod=1,solnbs=2:3)[,1]>8 |
 COS2(spaTab.MCA,mod=1,solnbs=2:3)[,1]>666]="."
> summary(as.factor(poilab)) # 306 OA highlighted
> plotmapbarq(Poly=OAmap,nclass=10,nvec=spaTab.MCA[[1]]$v[2, ],
      colrmp=colorRampPalette(Yl)(10),labels=poilab,cex=0.7,col="red",
      overlay=list(map=Wmap,border="black"))
> legend(locator(1),"vs2 48.65% ",bty="n")
```

and

```
> colCTRCOS2=rep("grey",2439)
> colCTRCOS2[poilab=="."]="red"
> plot(OAmap,border="white",col=colCTRCOS2)
```

3.4 FCA or MCA relative to a model

Classical correspondence analysis deals with decomposition of the chi-square of independence but in formula (1) $\hat{p}_{ij}$can be representing any other model than the independence model (Escofier 1984, Greenacre 2007). For spatial data an interesting application of this method is to analyse the deviations at a fine scale from the observed data or modelled data at a coarser scale. For the MoSes data the fine scale is at OA level and the coarser scale will be the same data aggregated at ward level.

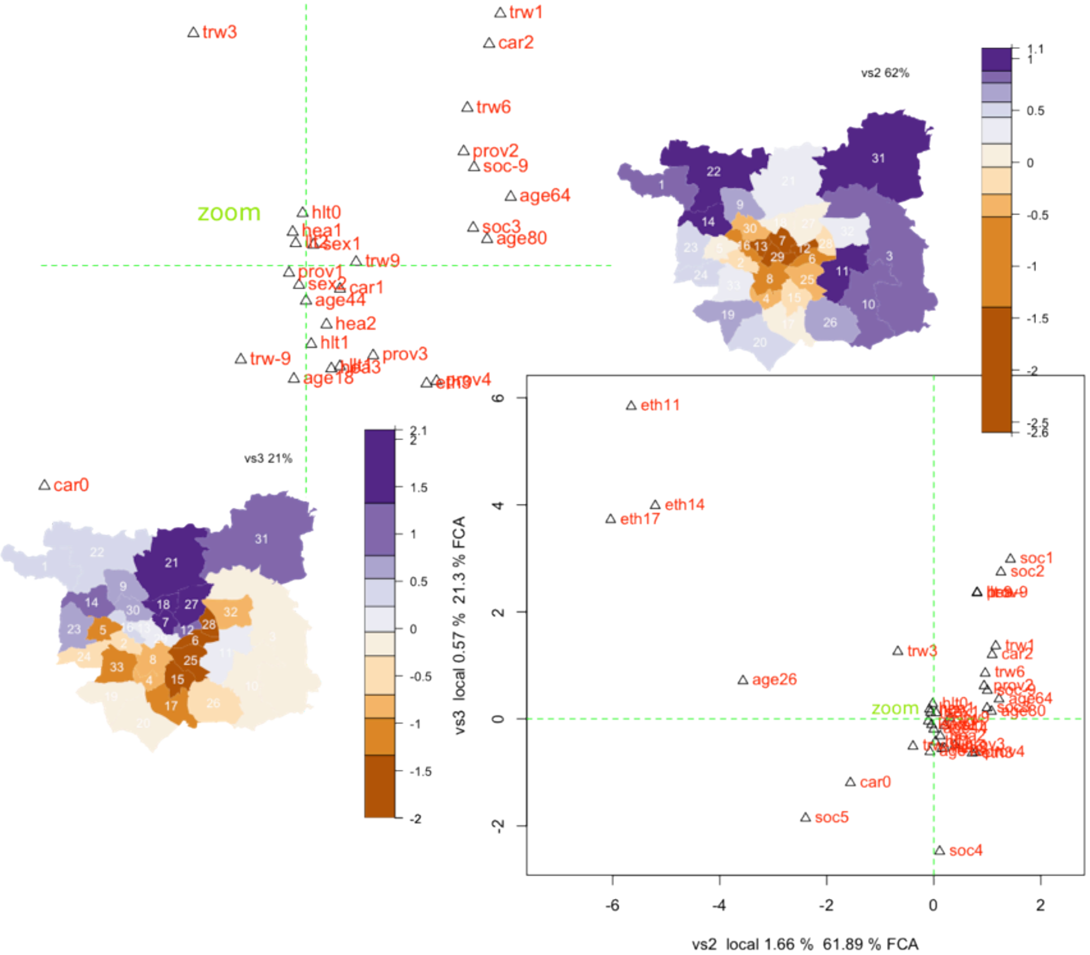

Figure 6: MCA of the chosen set of variables aggregated at Ward level- first two spatial components representing 83% of the lack of independence

The interpretations of the results from the MCA at Ward level are very similar to the one at the finer scale level OA (notice that, as the signs of the dimensions have changed both spatially and for the variables, the meaning is equivalent). A noticeable difference between the analysis at OA levels and at Ward levels is the presence of the so-called horseshoe effect or Guttmann effect on the `vs2` x `vs3` display, which expresses a relation like `vs3` vector related to the square of the `vs2` values, so that `vs3` associates the “extremes“, dissociated on `vs2`, in opposition to the “averages”.

When using the FCA/model analysis, if the data table and the model table have the same margins the chi-square distance keeps the same interpretation with respect to profiles distances on rows and columns (conditional probabilities) and no effects of the margins differences have to be taken into account when interpreting the results. This is the case when modelling the OA table using the Ward level values uniformly distributed in its OAs:

$$\hat{p}_{ij} = \left(\textstyle\sum_{a \in w(i)} n_{aj}\right)/noa_{w(i)}/N \qquad (2)$$

where $w(i)$ is the ward from which the i th OA belongs to and $noa_{w(i)}$ is the number of OA in that ward. The **R** code to build this model by creating the matrix, `Moa` is given below along with the call of the correspondence analysis of OAs in relation to this model being just the downscaled values of the Ward aggregation:

```
> Wnoa=summary(as.factor(WaOA[,1]))[c(1,12,23,28:33,2:11,13:22,24:27)]
> Mw=WardTab/Wnoa/sum(WardTab)
> Moa.w=spaTab #to initiate the dimensions and names
> all(WaOA[,2]==rownames(spaTab)) #TRUE
>     for (w in 1:33){
      Moa.w[WaOA[,1]==rep(paste("w",w,sep=""),2439),]=rep(1,Wnoa[w])%o%Mw[w,]
 }
> all(apply(Moa.w,2,sum)*sum(WardTab)==apply(WardTab,2,sum)) #TRUE
```

and the analysis is:

```
> spaTab.MCA.Moa.w <- FCA2(spaTab, E=Moa.w)
> summary(spaTab.MCA.Moa.w)
+++ FCA-   model(E=)   ++  2 modes+++
     ++ Contingency Table  spaTab   2439 39  ++
     -----Total Percent Rebuilt---- 100 %
     ++ Percent of lack of   model(E=)   rebuilt  ++  100 %
                    selected pctoafc >  0.5 %  total=  97.58512
    -no- --Sing Val-- --ssX-- --Global Pct--  --FCA--
vs1    1     0.733223 0.62165       86.48254 86.48254
vs2    2     0.191106 0.62165        5.87493  5.87493
vs3    3     0.104577 0.62165        1.75926  1.75926
vs4    4     0.094333 0.62165        1.43147  1.43147
vs5    5     0.072363 0.62165        0.84234  0.84234
vs6    6     0.064138 0.62165        0.66173  0.66173
vs7    7     0.057554 0.62165        0.53286  0.53286
++++               ++++
 Shown are selected  over  38  PT  with pct FCA > 0.5 %
```

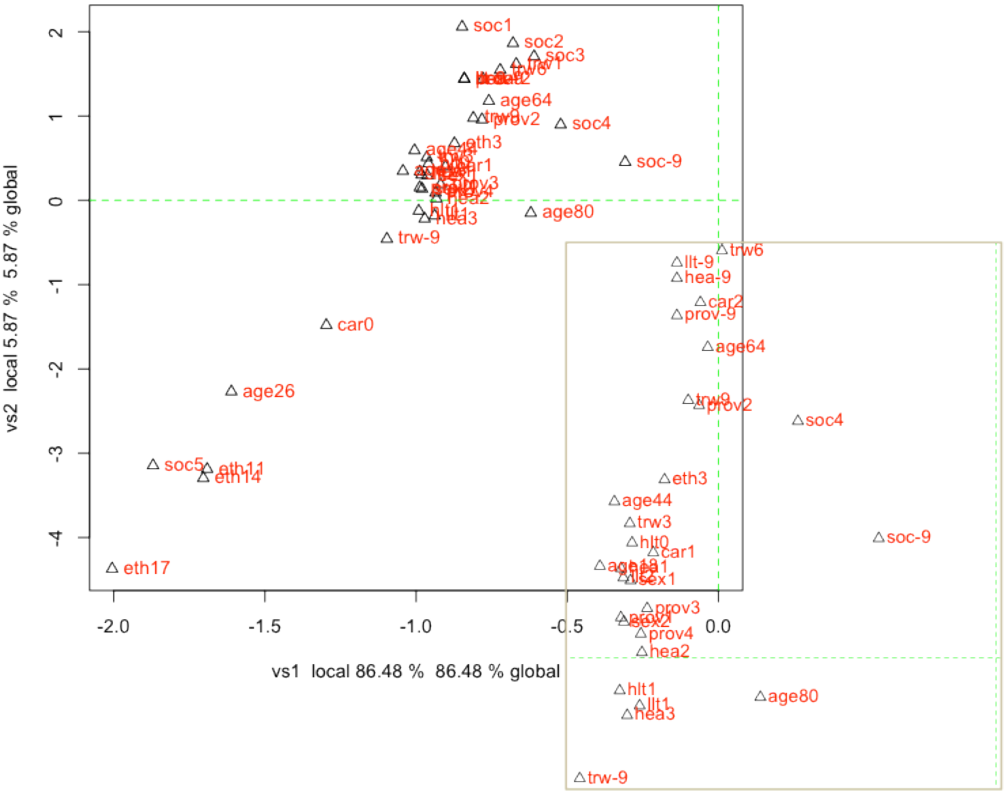

Figure 7: First two dimensions of the MCA in reference to a model: data are the chosen set of variables aggregated at OA levels, modelled by the ward level counts uniformly distributed (uniform downscaling)

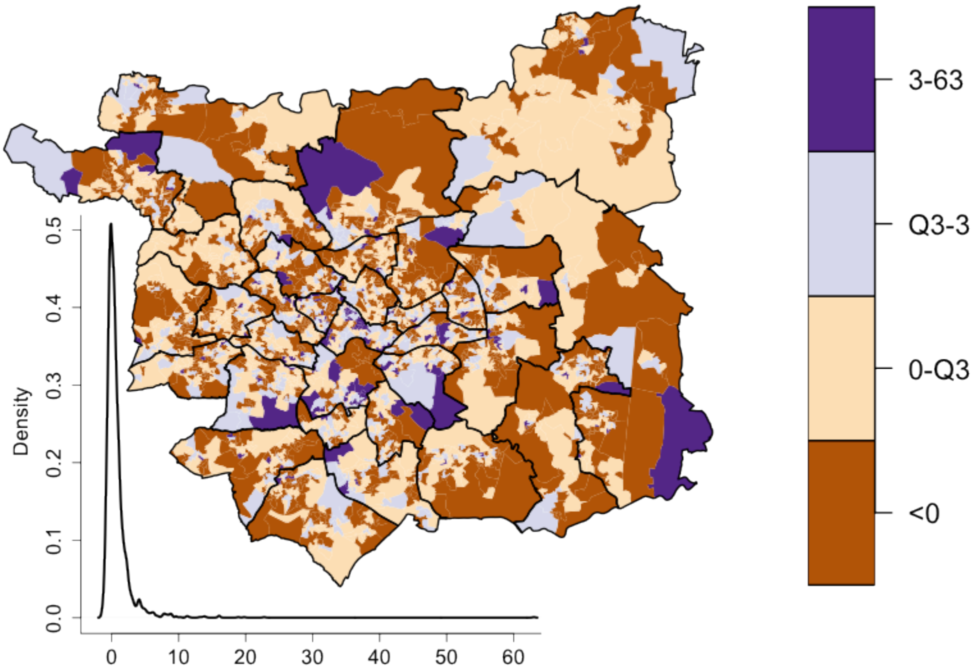

Figure 8: First spatial dimension (vs1, 86.5%) of the MCA in reference to a model OA/Ward. The values have been classified in 4 classes due to the very skewed distribution (lower panel), the legend of these classes refers to the third quartile Q3 = 1.001.

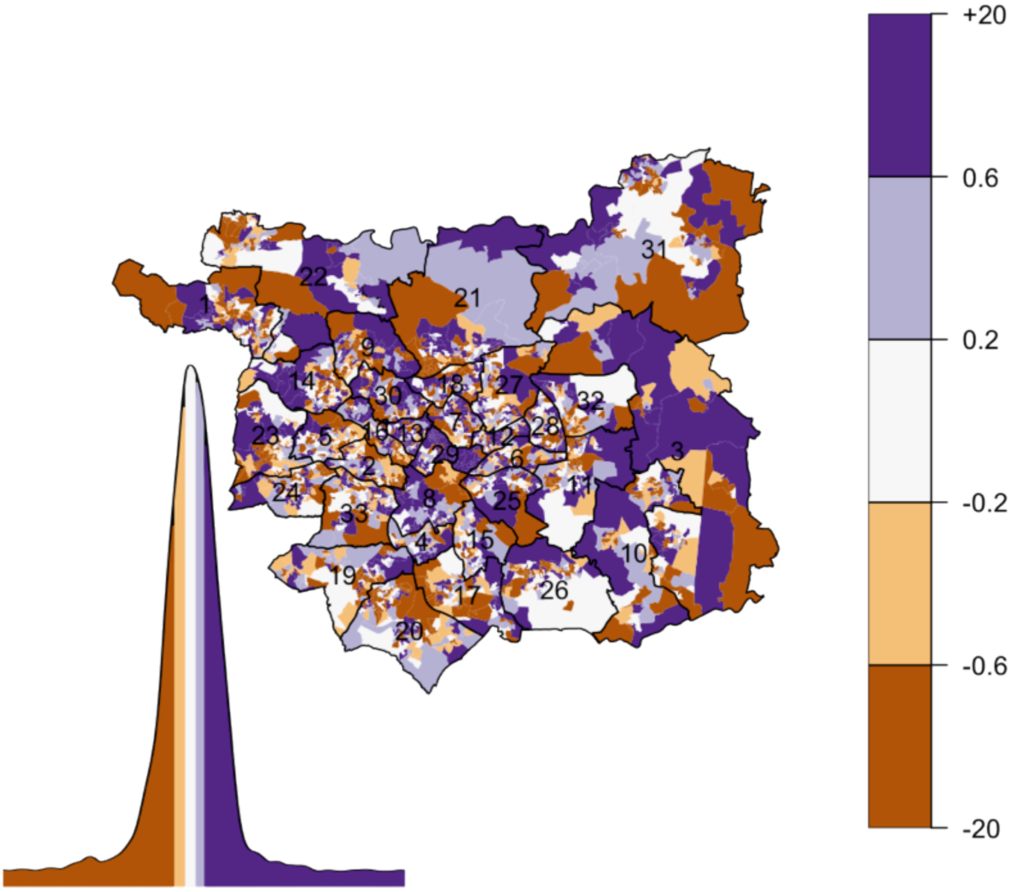

Figure 9: Second spatial dimension (`vs2`, 6%) of the MCA in reference to a model OA/Ward. The values have been classified in 5 classes due to the leptokurtic shape of the distribution, the legend representing the classes with their limits with an unscaled vertical bar.

This analysis is highlighting the OAs and variables categories irrespective to their Ward predictions (derived from aggregated observations), *i.e.,* departing from the Ward model. So it is not surprising that one get very spiky or skewed distributions of the values in Figure 8 and 9. Basically the OAs in dark blue and to a lesser extent in grey-blue are not following the profile of the categories we have seen in Figure 4 and 6. One can notice that these OAs (at least the ones in dark blue in Figure 8) are in the lower part of the distribution of the values in Figure 9 (dark brown) young (age 21-26) from non white origins (`eth11`, `eth14`, `eth17`) eth, in bad health, of social grade `soc5`, opposed to the more wealthy in good health with nonetheless some care weekly provided (provcar2: up to 19 hours/week) middle aged and older (up to 64 years of age) with a social grade `soc 1-3` but also `soc4` mainly from white origins (`eth3`).

4 Multiway Correspondence Analysis

So far a simple and multiple correspondence analysis have been demonstrated using **R** library **PTAk** and its respective functions but other **R** packages can be used to give the same results with similar functions: **ca**, **ade4** and others. In this section multiway correspondence analysis, which is part of the **PTAk R** package is going to be illustrated. Pursuing the arguments already mentioned in the introduction about the need of a multiway method extending the correspondence analysis methods, one can add or refine a few in the context of geosimulation. Concerning the spatio-temporal realm, sets of pairwise associations for some categories of the variables measured at different time may become difficult to interpret in the absence of simple gradients. So it may become easier to associate a temporal pattern to a variable pattern in the same way that in the previous sections a spatial pattern was associated to the latter. Therefore instead of looking at two sets of associated patterns one would like to get simultaneously three here (patterns in space, time and variable). In the same way one would prefer to separate observations done at different scales (Leibovici and Jackson 2011), nonetheless section 3.4 illustrated another way of dealing with different scales and integrating this aspect in the model itself.

4.1 a brief on the FCAk methodology

The multiplicative modelling aspect already existing in the PCA decomposition, *i.e.,* sum of products of vectors (each vector representing a pattern or a profile of the entry of the table) is extended to table with more than two entries in the PTAk decomposition. For 2 modes PCA or FCA are dealing with matrices, and for k>2 modes, PTAk or FCAk are dealing with multiway tables, that is arrays with more than two dimensions which can be also seen as tensors. From a mathematical algebraic point of view a tensor is a multi-linear operator: a 1 linear operator is a single vector, a bilinear operator correspond to a matrix, and tensors of order k>2, multi-linear operators, can be seen as multiple arrays. Elementary tensors, corresponding as well to the notion of linear operator of rank 1, are simply composed from the product (tensor product also termed outer product) of vectors. When performing a PCA or an SVD the first set of components looked for are, a vector capturing a pattern of the rows and a vector capturing a pattern of the columns, so that the bilinear operator made from the tensor product of these two vectors give the best approximation of the whole matrix analysed (best among the rank 1 bilinear operators):

$$r_1\,{}^t c_1 = r_1 \otimes c_1 = \underset{\substack{r\otimes c\ /\\ \|r\|=\|c\|=1}}{arg\min} \|M - r\otimes c\,\|^2 \qquad (3)$$

where r is a unit vector of the same dimension as the rows of the matrix M and *idem* for c as a vector of the same dimension as the columns of the matrix M. The expression $r_1 \otimes c_1$ is the called tensor product of the therein vectors and can be represented in a matrix form (left term of the equality) and as solution of the optimisation in (3) it is termed a principal tensor. The principal tensor optimisation in (3) extends to any dimensions, that is to a multiway table T:

$$r_1 \otimes c_1 \otimes d_1\ \ldots \otimes z_1 = \underset{\substack{r\otimes c\otimes d\ \ldots\otimes z\ /\\ \|r\|=\|c\|=\|d\|\ldots\|z\|=1}}{arg\min} \left\|T - r\otimes c\otimes d\ \ \ldots \otimes z\,\right\|^2 \qquad (4)$$

where $r, c, d \ldots, z$ are unit vectors of dimensions corresponding to the length of an entry in the multiway table or Array T. The norms, as in the expression $\|r\|$, in equation (3) and (4) are dependent of the choice of the metrics in the respective vector spaces: for example $\|r\|^2 = \sum_{ij} r_i d_{ij} r_j = {}^t r D r$ where D is a symmetric positive semi-definite matrix (*i.e.,* in order to make the norm positive or zero, and therefore the distances). The metric in the tensor space that is for matrices M or for tensor T is the tensor product of the metrics; see Leibovici (2010) for further details as well as how the algorithm carries on after the first principal tensor optimisation. To perform an FCAk where $k = 4$, that is a table T with 4 entries, we perform the PTAk of the multiway table $T_{rcdl} = \frac{p_{rcdl}}{p_{r...}p_{.c..}p_{..d.}p_{...l}}$ with the metrics D_R, D_C, D_D and D_L containing the margins of the table T along their respective entry, *e.g.,* $D_R = diag(p_{r...})$.

As with an FCA2, this analysis retrieves for the first set of components, also called from now the first principal tensor, the complete independence model, therefore gives also beyond this point in algorithm a decomposition of the chi-square statistic of independence like in the equation (1). Nonetheless, unlike an FCA2, it is necessary to perform an FCAk this way if one wants to fully decompose the chi-square. For example when $k = 3$, the chi-square can be expressed in a sum showing some parts involving two-way independences as well. Equation (5) gives some insights into this; see also Leibovici (2010):

$$\left(\frac{p_{rcd}}{p_{r..}p_{.c.}p_{..d}} - 1\right) = \left(\frac{p_{.cd}}{p_{.c.}p_{..d}} - 1\right) + \left(\frac{p_{r.d}}{p_{r..}p_{..d}} - 1\right) + \left(\frac{p_{rc.}}{p_{r..}p_{.c.}} - 1\right) + rsd \qquad (5)$$

where rsd is the what is left (the residual). Because this decomposition is orthogonal, that is, the decomposition stands when computing the squared norm of each tensor with the running elements being each part of (5), the FCAk, here an FCA3, is related to three independent FCA2s which are also independent from the decomposition of *rsd*.

4.3 a spatio-temporal FCAk

As a first example the same data structure as in 3.3 is used but with time added as a third mode, so an FCAk with k=3 modes is performed on the 6 periods of 5 years all put in an array of dimension `2439 x 7 x 39`:

```
> STOA=array(rep(0,2439*7*39),c(2439,7,39),
            dimnames=list(rownames(spaTab),
                  c("y01","y06","y11","y16","y21","y26","y31"),
                                             colnames(spaTab)))
> STOA[,1,]=spaTab;STOA[,2,]=buildspaTab(dat06);STOA[,3,]=buildspaTab(dat11);
      STOA[,4,]=buildspaTab(dat16);STOA[,5,]=buildspaTab(dat21);
      STOA[,6,]=buildspaTab(dat26);STOA[,7,]=buildspaTab(dat31
> dim(STOA)
[1] 2439    7   39
> STOA.FCAk <-FCAk(STOA)
>  summary(STOA.FCAk)
++++ FCA-  3 modes++++
     ++ Contingency Table  STOA   2439 7 39  ++
-----Total Percent Rebuilt---- 99.28517 %
     ++ Percent of lack of complete independence rebuilt  ++  93.49007 %
                       selected pctoafc >  0.2 %  total=  87.9884
                -no- --Sing Val--   --ssX-- --Global Pct--  --FCA--
vs111              1    1.000000 1.1233508      89.019388        NA
2439 vs111 7 39    3    0.054556 1.0032358       0.264958  2.41296
7 vs111 2439 39   10    0.218772 1.0754166       4.260568 38.80082
7 vs111 2439 39   11    0.098375 1.0754166       0.861495  7.84560
7 vs111 2439 39   12    0.072676 1.0754166       0.470189  4.28199
7 vs111 2439 39   13    0.049589 1.0754166       0.218902  1.99353
7 vs111 2439 39   14    0.048005 1.0754166       0.205145  1.86825
7 vs111 2439 39   15    0.045430 1.0754166       0.183723  1.67316
7 vs111 2439 39   16    0.040479 1.0754166       0.145864  1.32837
7 vs111 2439 39   17    0.035166 1.0754166       0.110088  1.00257
7 vs111 2439 39   18    0.030197 1.0754166       0.081173  0.73924
7 vs111 2439 39   19    0.022329 1.0754166       0.044385  0.40421
7 vs111 2439 39   20    0.016599 1.0754166       0.024526  0.22336
39 vs111 2439 7   49    0.137015 1.0260259       1.671183 15.21939
39 vs111 2439 7   50    0.044011 1.0260259       0.172426  1.57028
39 vs111 2439 7   51    0.038259 1.0260259       0.130306  1.18669
39 vs111 2439 7   52    0.036954 1.0260259       0.121563  1.10707
39 vs111 2439 7   53    0.035925 1.0260259       0.114886  1.04626
```

```
39 vs111 2439 7   54     0.034579 1.0260259       0.106442  0.96937
vs222             55     0.037645 0.0186726       0.126154  1.14888
7 vs222 2439 39   64     0.029085 0.0056972       0.075303  0.68578
7 vs222 2439 39   65     0.023334 0.0056972       0.048470  0.44142
7 vs222 2439 39   66     0.020774 0.0056972       0.038418  0.34987
7 vs222 2439 39   67     0.017669 0.0056972       0.027793  0.25311
7 vs222 2439 39   68     0.016465 0.0056972       0.024134  0.21978
39 vs222 2439 7  103     0.018917 0.0027728       0.031857  0.29012
39 vs222 2439 7  104     0.016813 0.0027728       0.025163  0.22915
39 vs222 2439 7  105     0.016437 0.0027728       0.024050  0.21902
vs333            109     0.018001 0.0115820       0.028847  0.26271
7 vs333 2439 39  118     0.016301 0.0026574       0.023655  0.21543

 ++++               ++++
 Shown are selected  over  121  PT  with pct AFC > 0.2 %
```

In the listing summary of the output from the FCAk, the name of the type of principal tensor is recognisable in the first column by the pattern of dimensions along with its main singular value name vs111, vs222 etc., then indicating the hierarchy of tensor solutions. For example the principal tensors with name type, 7 vs111 2439 39, correspond to the FCA2 mode once the 3-way table is collapsed on the time mode (of length 7). They are associated to the principal tensor vs111 *via* the time component, which is the same. These results show first of all that the complete independence represents 89% of the variability, so the margins encapsulate quite a lot of the spatio-temporal changes.

Now contracted in time (the tensors 7 vs111 2439 39), the correspondence analysis (equivalent to an FCA2 of the weighted sum over the years of the 2439 x 39 tables) captures 61% of the lack of complete independence, so not associated with 30 years of evolution. The results for this series of tensors, at least the main two first components, are very similar to the analysis just on year 2001 performed in the previous section.

Similarly 2.6% of the lack of complete independence is not due to OAs differences (the tensors 2439 vs111 7 39). In other words 2.6% of time-variable interaction is independent of the OAs differences, and finally 21.1% of the spatio-temporal variation is not due to variables profiles (the tensors 39 vs111 2439 7); this is population variation only. Figure 10 represents the main captured effect (vs n°49), which is correlated with the period differences; this is linked to a year component expressing a relatively linear gradient. The wards on the edges of the district in North or on the East along with wards n°19 and n°26 but also some wards close the north of the city centre (ward n°8) are increasing their population whilst areas immediately surrounding round these as well as east and west of the city centre show a diminution in population.

So a total of nearly 85% of the lack of complete independence can be attributed to 2-way interaction, leaving nonetheless altogether 15% of pure 3-way interaction that cannot be analysed by an MCA.

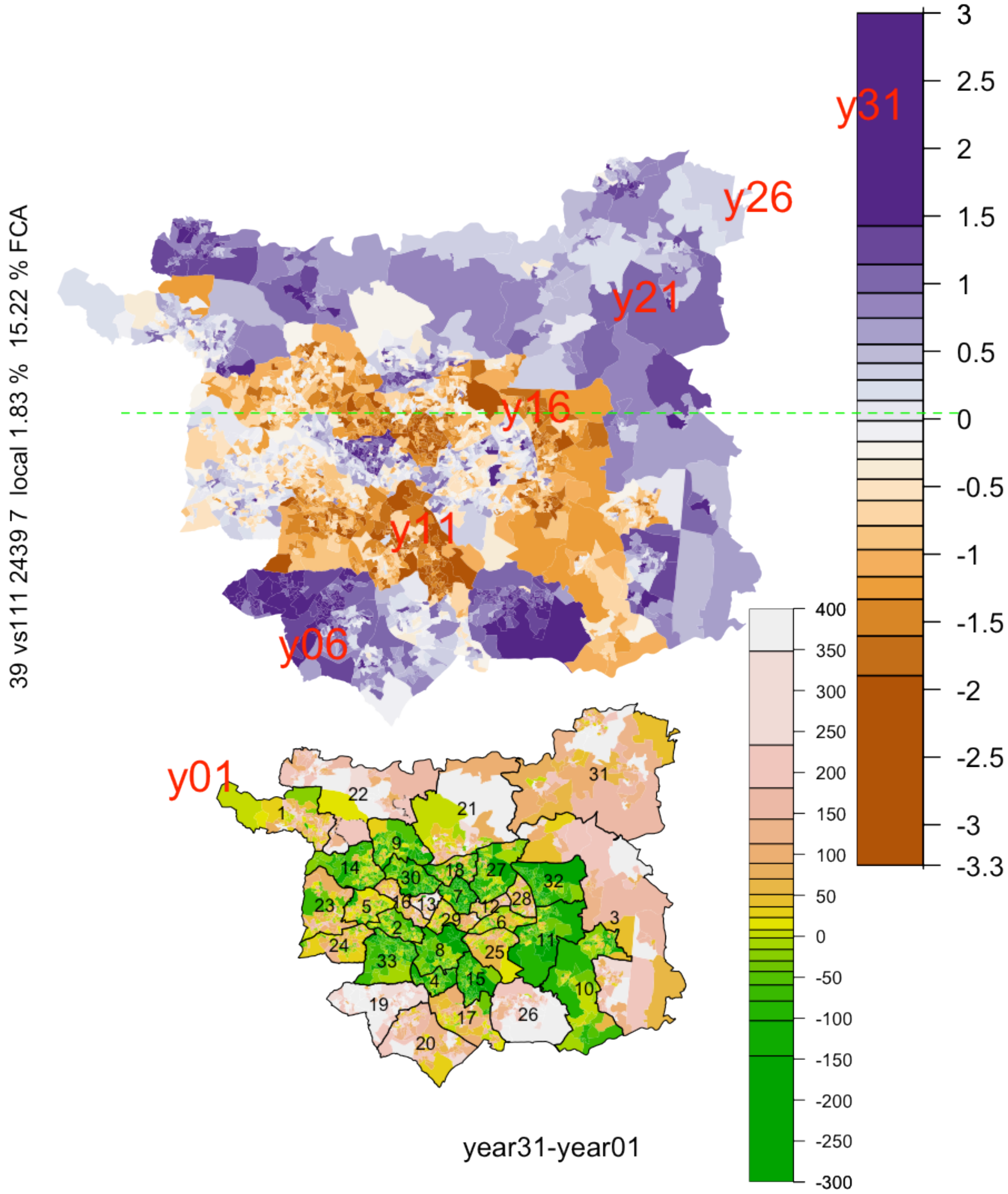

Figure 10: FCAk at OA levels of the Contingency Table: STOA 2439 7 39; Spatial-temporal evolution of the population as expressed by the tensor vs n°49 expressing 15% of the lack of complete lack independence but associated with a marginal effect along the variable mode. - bottom panel: spatial variation of the difference of population counts along the period of 30 years -

Using a normalised spatial entropy index, based on co-occurrences counts within a given collocation distance (Leibovici and Birkin 2013), applied to the categorical variable representing the legend of the map (the spatial tensor mode representation) one gets a range of 0.45 to 0.52 according to a collocation variation from 2000m to 5000m. The normalised index is 1 when the distribution of co-occurrences is uniform, so this indicates here a fairly strong spatial structure.

The linear gradient in time nonetheless also exists within the three-way interaction principal tensors, Figure 11 showing vs222 and its first associated principal tensor along the time mode, with again similar pattern of the variables. Figure 12 shows for each of the variable profiles a linear time trend added to the similar main "constant" effects from principal tensor n°10 and n°11. Notice that as the time mode component is the same for vs222 and any 7 vs222 2439 39 tensors (see Leibovici 2010), the 7 years appears on the first diagonal of the plot in Figure 11.

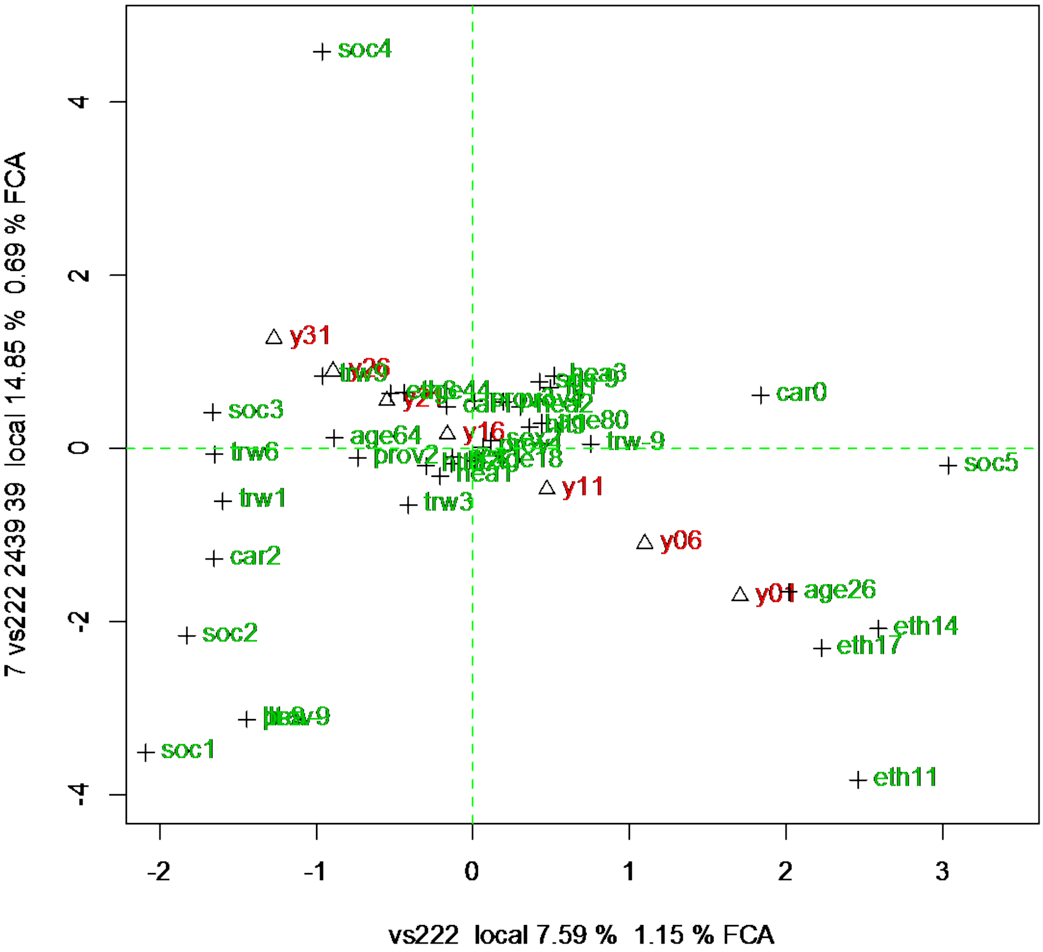

Figure 11: FCAk at OA levels of the Contingency Table: STOA 2439 7 39; tensor vs n°55 and n°64 expressing altogether 1.84% of the lack of complete lack independence each as a three-way interaction (spatial-temporal-variables)

Therefore the opposition already seen in 2-modes analysis between soc5 and the other social grades is associated to a change over the years in the distribution of social grades: soc5 more prominent at the beginning of the period in less towards the end for some OAs.

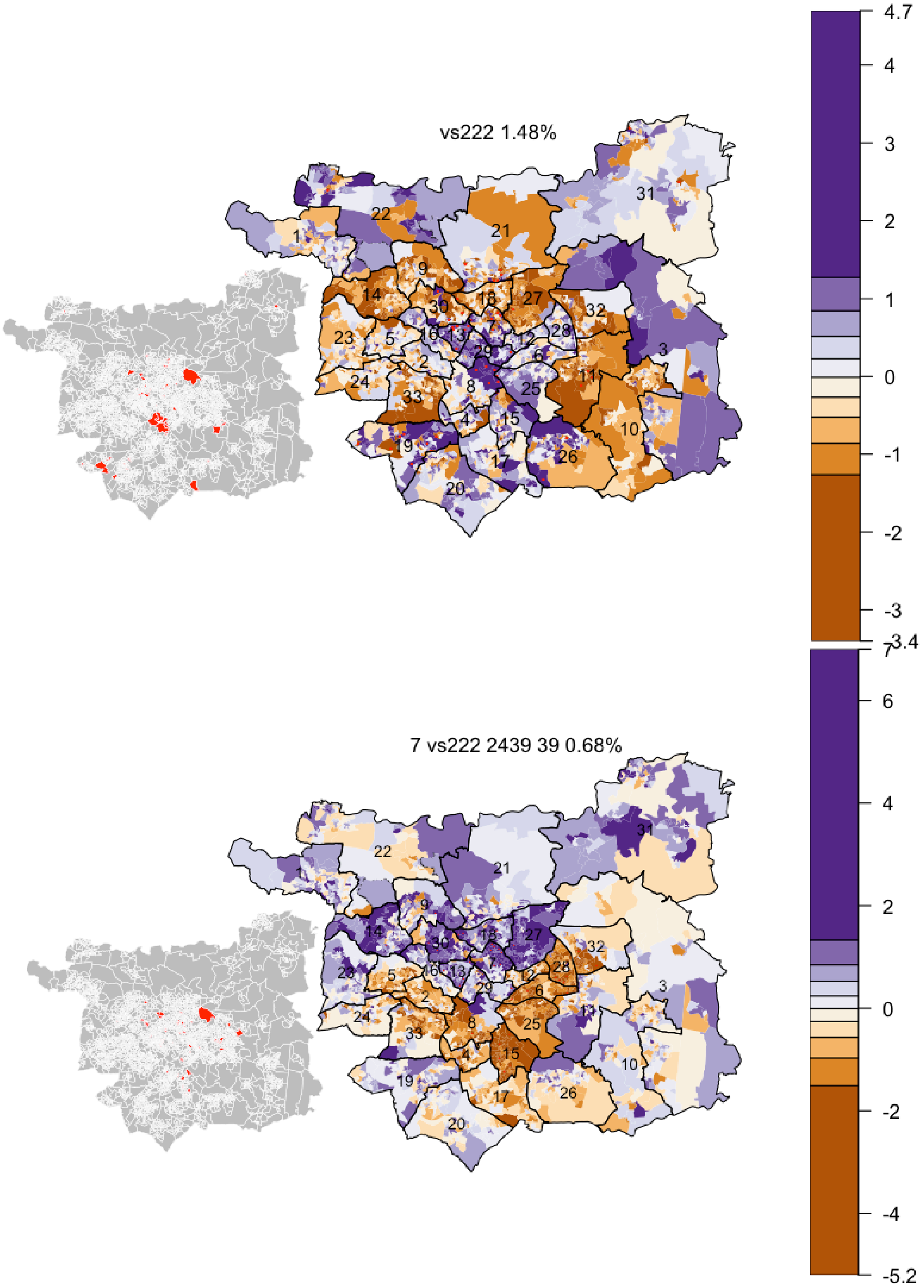

Figure 12: FCAk at OA levels of the 7 Contingency Tables over the 30 years period: the data array STOA 2439 7 39; spatial components for principal tensors vs n°55 and n°64, with a thumbnail the OAs most contributing or best representing, highest CTR or COS2, onto the principal tensor: for vs222 (n°55) 47 OAs highlighted, for the other principal tensor 42 OAs highlighted (the large single OA northern central is within the ward 27- Roundhay).

The spatial structure of Figure 12 exhibits some pockets of strong differences either on the positive or negative side of the dimensions with nonetheless expressing the extremes of regular circular gradients on both spatial principal tensors. Rings of structure in the inner suburb (for example with the wards 4, 33, 2, 5, 14, 9, 30, 18, 27, 29 and 11) as opposed to the city centre can be depicted. The bottom map has also a north south differentiation for positive and negative values.

4.4 spatial scale analysis with FCAk

Being able to consider adding more of the data structure and sampling design directly into the data to be analysed increases the possibilities of analyses. The important aspect of the choice of the scaless, combined with the possibility of interaction through scale can be analysed using an FCAk of the table augmented with the scale aggregation aspects (Leibovici and Jackson 2011), and, like in section 3.4 a correspondence relative to a model whilst using FCAk can be also performed, or both. For the MoSes data example the scale factor is the hierarchy of census/administrative boundaries of the Leeds district area: OAs (2439 units), LSOAs (476), MSOAs (108) and Wards (33) based on the census 2001 boundaries.

4.4.1 aggregating the data and downscaling

Below is the **R** code to be able to aggregate the frequencies at different scale levels and to downscale these aggregated values, so to illustrate and uniform homogeneous model coming a coarser scale.

```
> CensusPopOA <- read.delim("OA.SOA.LSOA.Pop2001.txt")
> nlevels(as.factor(CensusPopOA$OA_code))
 [1] 2439
>  nlevels(as.factor(CensusPopOA$LSOA_code))
 [1] 476
>  nlevels(as.factor(CensusPopOA$MSOA_code))
 [1] 108
>  all(rownames(WaOA)==rownames(CensusPopOA))
[1] TRUE
>  LSoa=as.factor(CensusPopOA$LSOA_code)
#####
> aggTab3 <-function(Tab3=STOA,by1=levels(LSoa),repby1=CensusPopOA$LSOA_code){
 #repby1 contains the levels and has the same length as dimanmes(STOA)[[1]]
 #by1 is a grouping factor
out=array(rep(0,length(by1)*dim(Tab3)[2]*dim(Tab3)[3]),c(length(by1),dim(Tab3)[2],dim(
 Tab3)[3]))
            dimnames(out)[[1]]=by1
            dimnames(out)[[2]]=dimnames(Tab3)[[2]]
            dimnames(out)[[3]]=dimnames(Tab3)[[3]]
 for (i in 1:length(by1)){
      out[i,,]=apply(Tab3[repby1==dimnames(out)[[1]][i],,],c(2,3),sum)
 }
 return(out)
}##
> LSOATab=aggTab3()
> MSOATab=aggTab3(by1=levels(as.factor(CensusPopOA$MSOA_code)),
      repby1=CensusPopOA$MSOA_code)
> WTab=aggTab3(by1=levels(as.factor(WaOA[,1])),repby1=WaOA[,1])
#####
downScale3 <-function(fine=STOA,agg=WTab,repby1=WaOA[,1]){
      noa=summary(as.factor(repby1),maxsum=dim(agg)[1])
      Mass=agg/noa
      out=fine
      for (w in 1:dim(agg)[1]){
      out[repby1==rep(dimnames(agg)[[1]][w],dim(fine)[1]),,]=rep(1,noa[w])%o%Mass[w,,]
 }
 return(out)
}##
> MOA.Ward=downScale3()
> MOA.Msoa=downScale3(,agg=MSOATab,repby1=CensusPopOA$MSOA_code)
> MOA.Lsoa=downScale3(,agg=LSOATab,repby1=CensusPopOA$LSOA_code)
```

4.4.2 multiscale FCAk

The downscaled observations representing different scales of measuring an event can then be analysed simulataneously in order to detect effects across or particular ot specific scales.

```
> STOAscales= array(rep(0,2439*7*39*4),c(2439,7,39,4))
> STOAscales[,,,1]=STOA
> STOAscales[,,,2]=MOA.Lsoa
> STOAscales[,,,3]=MOA.Msoa
> STOAscales[,,,4]=MOA.Ward
> dimnames(STOAscales)[c(1,2,3)]=dimnames(STOA)
> dimnames(STOAscales)[[4]]=c("OA","LSOA","MSOA","Ward")
> STOAscales.FCAk <- FCAk(STOAscales,addedcomment=" 4 upscaled aggregated ")
> summary(STOAscales.FCAk,testvar=0.2)
 +++ FCA-  complete independence  ++  4 modes+++
     ++ Contingency Table  STOAscales   2439 7 39 4  ++
     4 upscaled aggregated
     -----Total Percent Rebuilt---- 99.76836 %
     ++ Percent of lack of  complete independence  rebuilt  ++  98.28229 %
                    selected pctoafc >  0.2 %  total=  92.35749
                  -no- --Sing Val--   --ssX-- --Global Pct--  --FCA--
vs1111              1     1.000000 1.1558739      86.514627       NA
2439-4 vs111 7 39  12     0.054556 1.0032358       0.257503  1.90950
7-39 vs111 2439 4  25     0.239979 1.0812272       4.982352 36.94634
7-39 vs111 2439 4  26     0.134887 1.0812272       1.574090 11.67257
7-39 vs111 2439 4  27     0.073776 1.0812272       0.470895  3.49190
7-4 vs111 2439 39  29     0.124506 1.0246931       1.341126  9.94504
7-4 vs111 2439 39  30     0.054296 1.0246931       0.255051  1.89131
7-4 vs111 2439 39  31     0.048429 1.0246931       0.202909  1.50466
7-4 vs111 2439 39  32     0.039944 1.0246931       0.138035  1.02359
7-4 vs111 2439 39  33     0.029114 1.0246931       0.073330  0.54378
7-4 vs111 2439 39  34     0.018695 1.0246931       0.030238  0.22423
7-vs222            67     0.100183 0.0194099       0.868315  6.43894
7-39 vs222 2439 4  73     0.059012 0.0145015       0.301285  2.23416
7-39 vs222 2439 4  74     0.031344 0.0145015       0.084996  0.63028
7-4 vs222 2439 39  77     0.032669 0.0135770       0.092336  0.68471
7-4 vs222 2439 39  78     0.027162 0.0135770       0.063831  0.47333
7-4 vs222 2439 39  79     0.018610 0.0135770       0.029962  0.22218
7-vs333           115     0.019666 0.0013669       0.033460  0.24812
39-4 vs111 2439 7 173     0.123111 1.0162036       1.311233  9.72337
39-4 vs111 2439 7 174     0.021623 1.0162036       0.040450  0.29996
39-vs222          179     0.038930 0.0035632       0.131115  0.97228
39-7 vs222 2439 4 185     0.020725 0.0022243       0.037161  0.27556
4-vs222           250     0.025878 0.0038805       0.057937  0.42963
4-7 vs222 2439 39 259     0.024015 0.0021558       0.049894  0.36998
vs2222            358     0.017747 0.0036607       0.027249  0.20207

 ++++                ++++
 Shown are selected  over  367  PT  with pct FCA > 0.2 %
```

The independence, that is only marginal effects, takes 86.5% of the variability, leaving 13.5% explained by lack of independence between the modes; about half of the lack of complete independence (37%+11.5%+3.5% from principal tensors 25, 26 and 27) is due to OAs and scale interaction with marginal effect from time and variable, whilst around 15% is due to spatial variation of the variables marginally in time and scale (principal tensors 29-34).

Another 10% (tensors 173-174) expresses two-way interactions between space and time marginally t scale and variable modes, which leaves 25% some of the lack of complete independence as due to multiway interactions with 10% (tensors 67-79) associated to marginal time effect. So, 13.5% of this lack of complete independence (15% left minus % due to tensors 179-259) is due to multiway interactions without initial marginal effect involved.

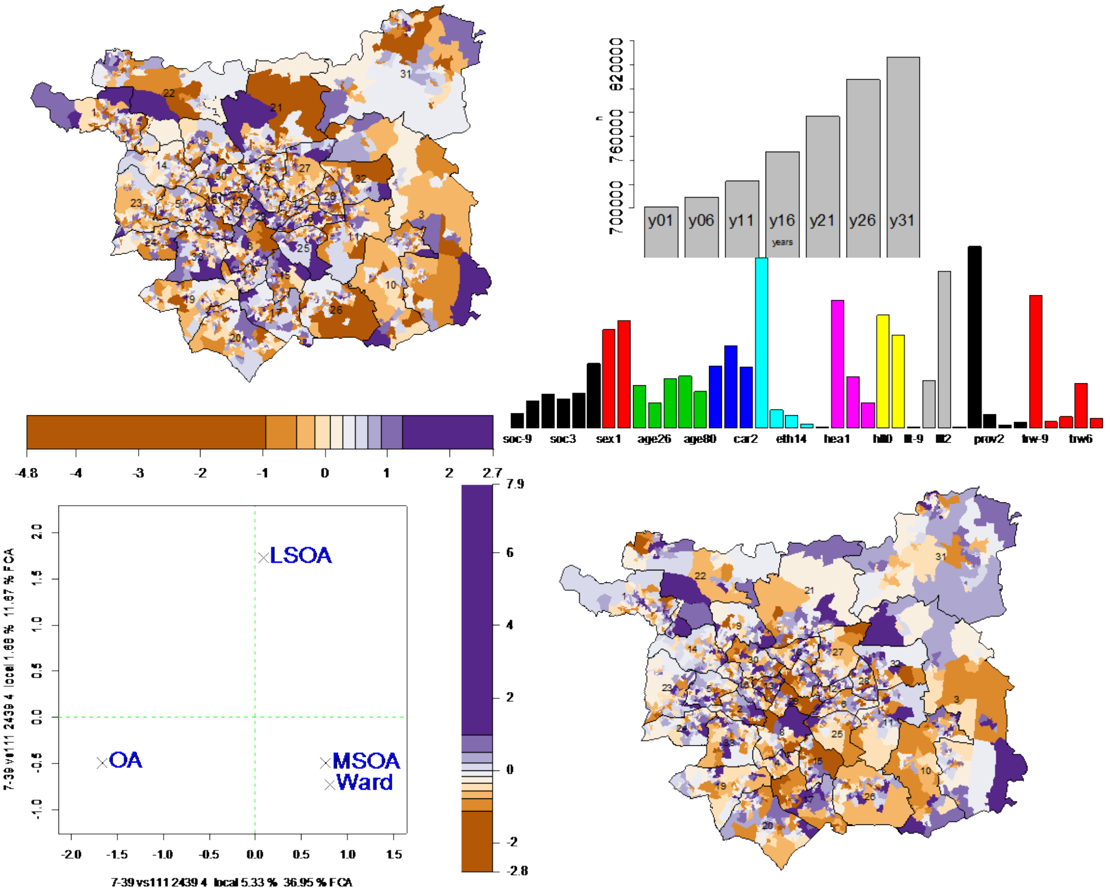

Figure 13: FCAk with a mode as multiple scales (downscaled uniformly at OA levels), first two principal tensors associated to marginal effect for year and variable mode – equivalent to an FCA2 space x scale.

In Figure 13, the spatial –scale interaction is shown representing 48.5% of the lack of independence. IT shows that the Ward and MSOA variations are similar and not in agreement with OA variations for the OAs with extreme values on the spatial component map of the axis (top map in Figure 13); some the units are the same on the other map associated with an opposition of LSOA to the other scale levels.

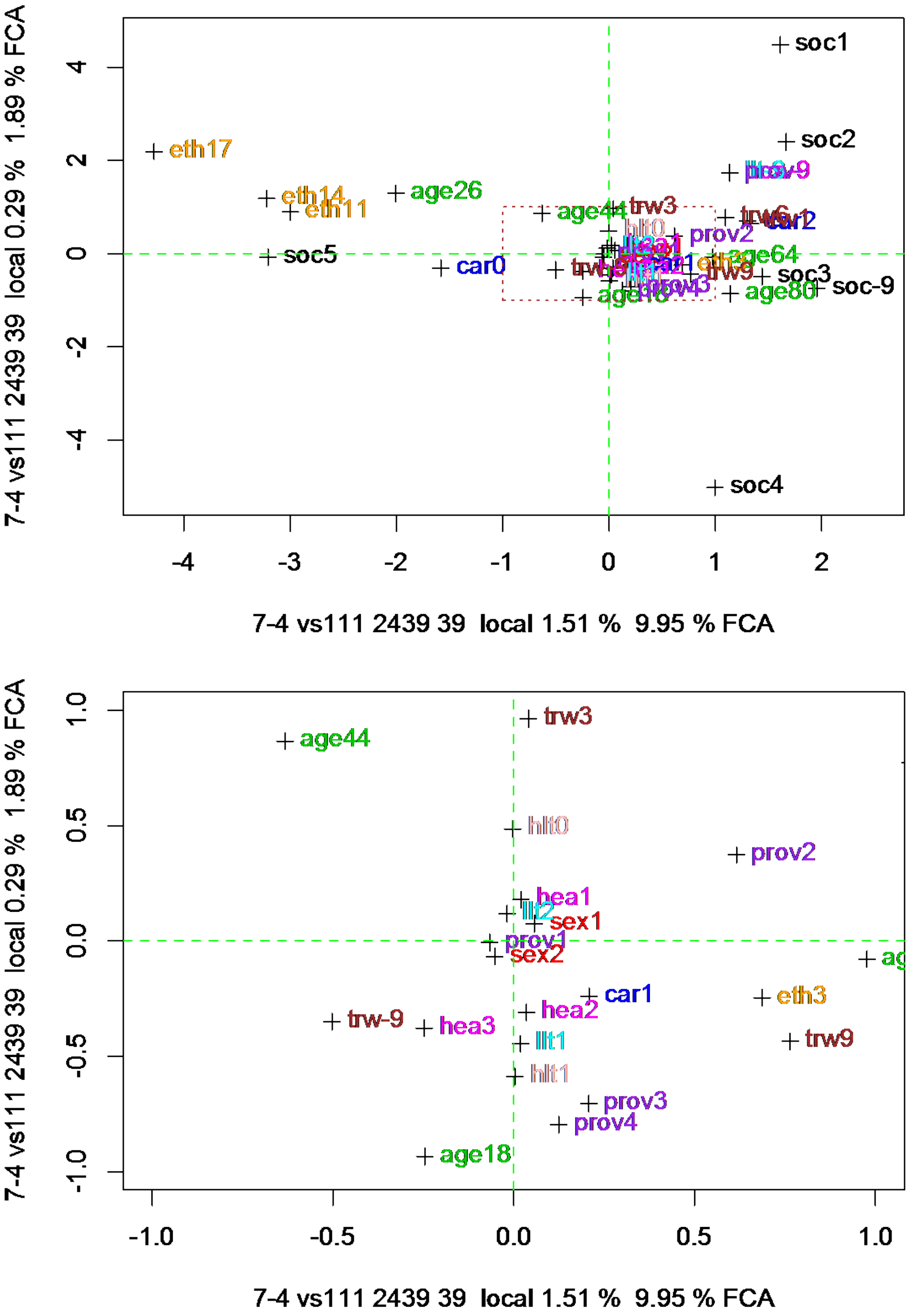

Figure 14: FCAk with a mode as multiple scales (downscaled uniformly at OA levels), first two principal tensors associated to marginal effect for year and scale mode – equivalent to an FCA2 space x variable (the bottom plot is a zoom of the dotted brown rectangle from the top plot)

Marginally associated to time and scale, Figure 14 shows again on the horizontal axis (see also Figure 15 top panel) a strong association ethnicity and social grades, people in social category soc5 (*hrscogrd* E that is on benefit or unemployed) being more often non-Caucasian white and of younger age (18-26 years old). The middle aged to the eldest (44 years old and over) had more likely, up to 19 hours of care provided per week, as opposed to the younger (18-44 years old) but the youngest (<18 years) had most of the cares (prov3 and prov4, more than 20 hours per week), this last aspect irrespective of ethnicity but associated to a manual workers social categories (see also Figure 15 bottom map).

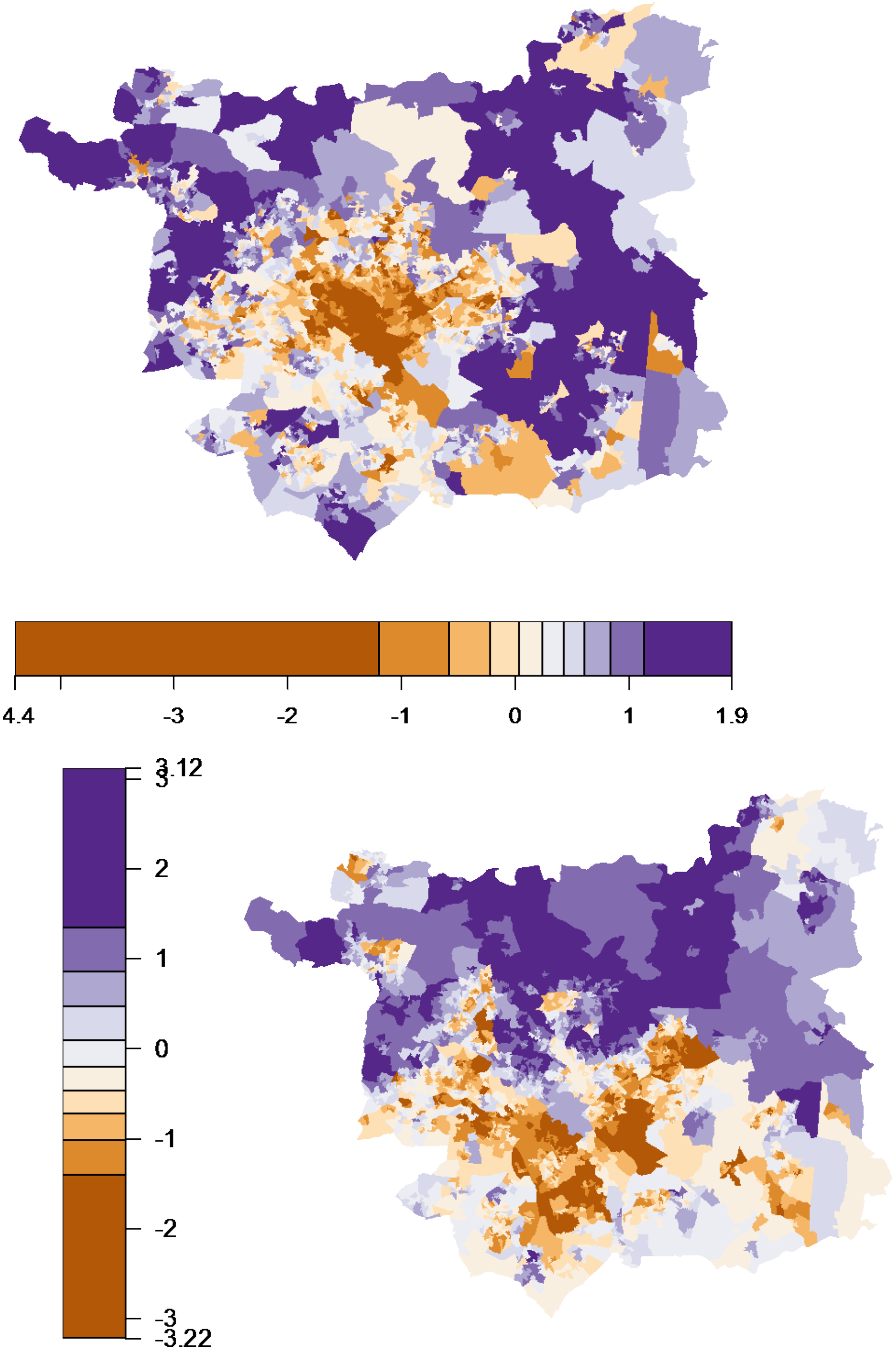

Figure 15: see Figure 14 caption, corresponding spatial components (top is the horizontal axis of Figure 14)

Figure 15 shows an apparent spatial pattern associated with the above description of the variable effects: Figure 15 (top) opposes the centre, non white Caucasian identified above, to the outskirt of the district and more the north-west and a large area at the East, (bottom) shows a north south divide (younger manual workers in the south with more care provided to).

4.4.3 FCAk in relation to a model

With the FCAk it is also possible to perform a correspondence relative to a model. Here an example where the observed proportions are modelled by the aggregated values at Ward level. As similar to analysing the residuals from a model, the type of analysis, here, is highlighting OAs which do not follow the Ward aggregated model. In Figure 16 (bottom panel) can be seen the very few (large spread of the percentile legend) OAs which departs strongly from that model (in brown). The variable component associated to this “outliers” pattern is not very different from was described before (here one dimension) indicating that the departure from the model is more “quantitative” than “qualitative”.

```
> STOA.FCAk.MOA.Ward <- FCAk(STOA,E=MOA.Ward/sum(MOA.Ward))
> summary(STOA.FCAk.MOA.Ward,testvar=0.1)
 +++ FCA-    model(E=)    ++   3 modes+++
     ++ Contingency Table  STOA    2439 7 39   ++
     -----Total Percent Rebuilt---- 98.57719 %
     ++ Percent of lack of    model(E=)    rebuilt   ++  98.57719 %
                    selected pctoafc >  0.1 %  total=  96.85047
                -no- --Sing Val--    --ssX-- --Global Pct--   --FCA--
vs111              1     0.695282 0.5695688       84.87425 84.87425
2439 vs111 7 39    3     0.035160 0.4847447        0.21705  0.21705
7 vs111 2439 39   10     0.163823 0.5325014        4.71196  4.71196
7 vs111 2439 39   11     0.085116 0.5325014        1.27196  1.27196
7 vs111 2439 39   12     0.066459 0.5325014        0.77547  0.77547
7 vs111 2439 39   13     0.047314 0.5325014        0.39304  0.39304
7 vs111 2439 39   14     0.043218 0.5325014        0.32793  0.32793
7 vs111 2439 39   15     0.040730 0.5325014        0.29127  0.29127
7 vs111 2439 39   16     0.037365 0.5325014        0.24512  0.24512
7 vs111 2439 39   17     0.031459 0.5325014        0.17376  0.17376
7 vs111 2439 39   18     0.028583 0.5325014        0.14344  0.14344
39 vs111 2439 7   49     0.104617 0.5008396        1.92157  1.92157
39 vs111 2439 7   50     0.041012 0.5008396        0.29531  0.29531
39 vs111 2439 7   51     0.036322 0.5008396        0.23163  0.23163
39 vs111 2439 7   52     0.035073 0.5008396        0.21598  0.21598
39 vs111 2439 7   53     0.034018 0.5008396        0.20318  0.20318
39 vs111 2439 7   54     0.033001 0.5008396        0.19121  0.19121
vs222             55     0.038515 0.0183176        0.26044  0.26044
7 vs222 2439 39   64     0.024561 0.0050178        0.10591  0.10591

 ++++               ++++
 Shown are selected  over   121   PT   with pct FCA > 0.1 %
```

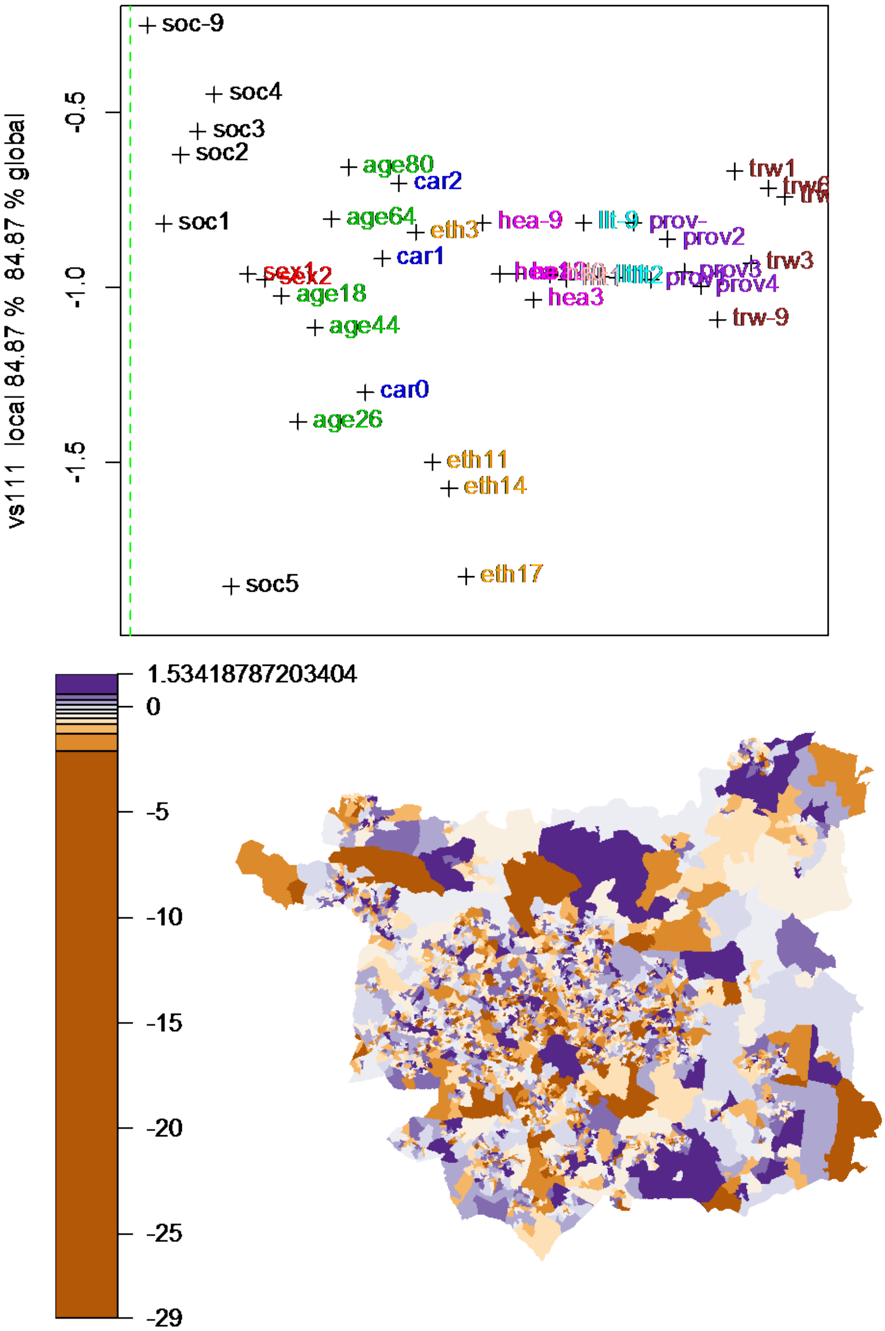

Figure 16: First principal tensor of the FCAk in relation to the ward model expressing 85% of the departure from that model. (top: the horizontal spread is artificial as only one dimension is displayed). The time component is expressing an average over the 7 years, so is not displayed.

Notice the strong extremum effect illustrated by the quantile legend of the spatial map of figure 16; this expresses very few but quite extreme negative values (in brown), dispersed mainly far from the inner city and associated with the presence of few (as a not the generality for these areas) of young (18 to 26 years old: age26) Asian, Caribbean, and African ethnics (eth11, eth14, eth17) on benefit or unemployed (soc5).

4.4.4 FCAk for domain interactions

This approach is similar to the Burt table analysis but in using a multiway context. So far, we have analysed multiway interactions between very different modes: space, time, variables, but with in the variables different domains can be “separated”: health-realted with *hea, lti , hlt and pro*, socio-demographic with *age, sex, soc , eth* and *car*. The variable “transport to work” can be health related but also socio-demographic related and was not used here to simplify.

```
> spaDom01.FCAk=FCAk(spaDom01)
> summary(spaDom01.FCAk)
+++ FCA-  complete independence  ++  3 modes+++
     ++ Contingency Table  spaDom01   2439 20 14  ++
     -----Total Percent Rebuilt---- 98.94027 %
     ++ Percent of lack of  complete independence  rebuilt  ++  95.20483 %
                    selected pctoafc >  0.5 %  total=  86.7174
                 -no- --Sing Val--   --ssX-- --Global Pct--   --FCA--
vs111               1     1.000000 1.2836955      77.90010        NA
2439 vs111 20 14    3     0.189772 1.0420572       2.80545 12.69441
2439 vs111 20 14    4     0.054540 1.0420572       0.23173  1.04854
2439 vs111 20 14    5     0.046467 1.0420572       0.16820  0.76110
20 vs111 2439 14   17     0.101154 1.0206284       0.79708  3.60673
20 vs111 2439 14   18     0.064391 1.0206284       0.32299  1.46148
20 vs111 2439 14   19     0.043422 1.0206284       0.14688  0.66462
14 vs111 2439 20   31     0.302685 1.1780322       7.13706 32.29453
14 vs111 2439 20   32     0.162238 1.1780322       2.05043  9.27801
14 vs111 2439 20   33     0.126284 1.1780322       1.24232  5.62137
14 vs111 2439 20   34     0.097632 1.1780322       0.74254  3.35991
14 vs111 2439 20   35     0.093354 1.1780322       0.67889  3.07193
14 vs111 2439 20   36     0.081693 1.1780322       0.51988  2.35242
14 vs111 2439 20   37     0.074683 1.1780322       0.43450  1.96605
14 vs111 2439 20   38     0.062426 1.1780322       0.30358  1.37368
14 vs111 2439 20   39     0.047397 1.1780322       0.17500  0.79185
14 vs111 2439 20   40     0.044972 1.1780322       0.15755  0.71289
14 vs111 2439 20   41     0.041032 1.1780322       0.13115  0.59345
vs222              50     0.066577 0.0429777       0.34529  1.56239
14 vs222 2439 20   80     0.051006 0.0164032       0.20266  0.91704
14 vs222 2439 20   81     0.042127 0.0164032       0.13825  0.62555
vs333              99     0.062754 0.0239564       0.30678  1.38815
14 vs333 2439 20  129     0.040258 0.0093509       0.12625  0.57129

 ++++               ++++
 Shown are selected  over  91  PT  with pct FCA > 0.5 %
```

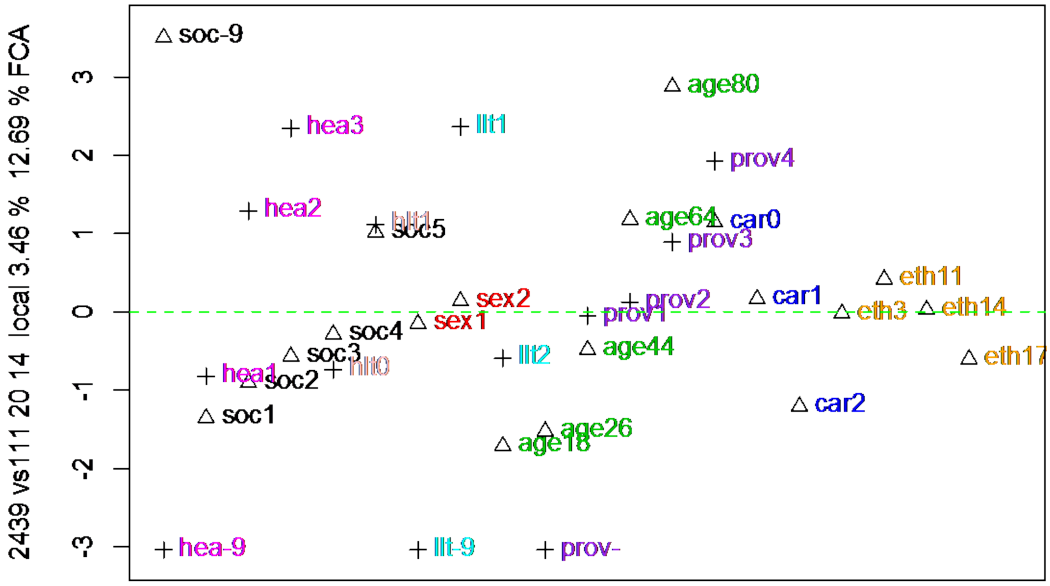

Figure 17: Principal tensor n°3 associated to marginal spatial: overall (for all OAs) domain interaction (socio-demographic and health related) in 2001

Among the 22% of variability that are linked to the lack of independence between the domains of spatial, health and socio-economic description of the data, 12% of it reflects an overall spatial effect (average spatial effect) with associations, on the positive side of the axis in Figure 17, of old people, without a car, potentially on benefit or unemployed, not in a good health or in a fairly good health needing a lot of health care and in a long term illness, as opposed to relatively young, with more than 1 car, professionals an non-manual workers, in good health with no need of care. Notice a tendency for this latter to happen for white and black African more than the reverse, whilst other ethnic groups and particularly the British or Irish or other white can be "equally" in each side.

Nonetheless and independently to the health domain the most part of these 22% of lack of independence, are to be found expressed within the tensor n°31 (Figure 18) representing 32% of these 22% and showing a strong association as we have seen before of Asian and African ethnicity along with on benefits or unemployed, young with no car in the city centre (brown zones) and on the outskirt of the city in few places towards the south east of the district. The north part of the district (in dark blue) pointing out more wealth among the elder and eldest of the population more likely from British or Irish or other white origins.

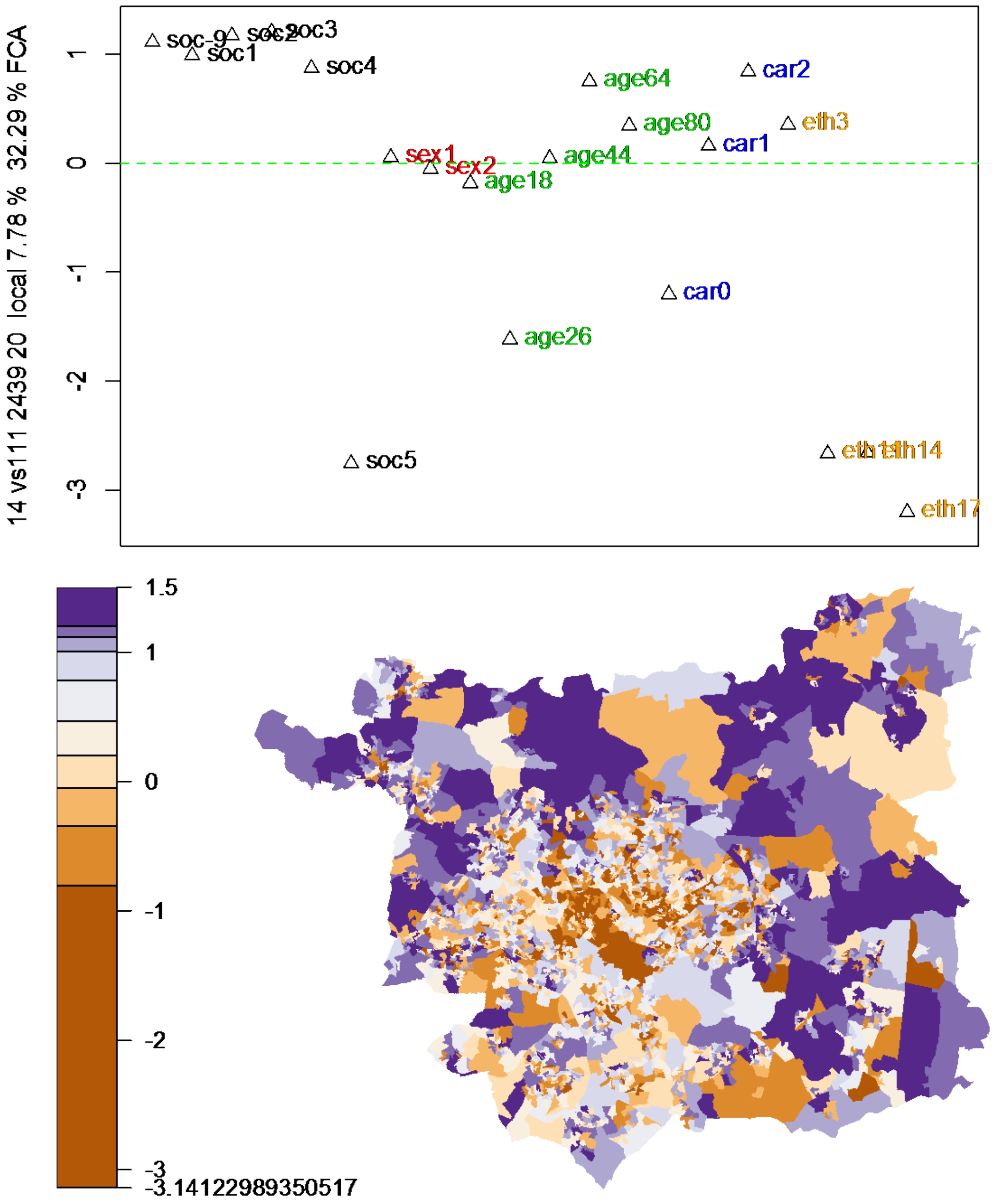

Figure 18: Principal Tensor n°31 of the FCAk

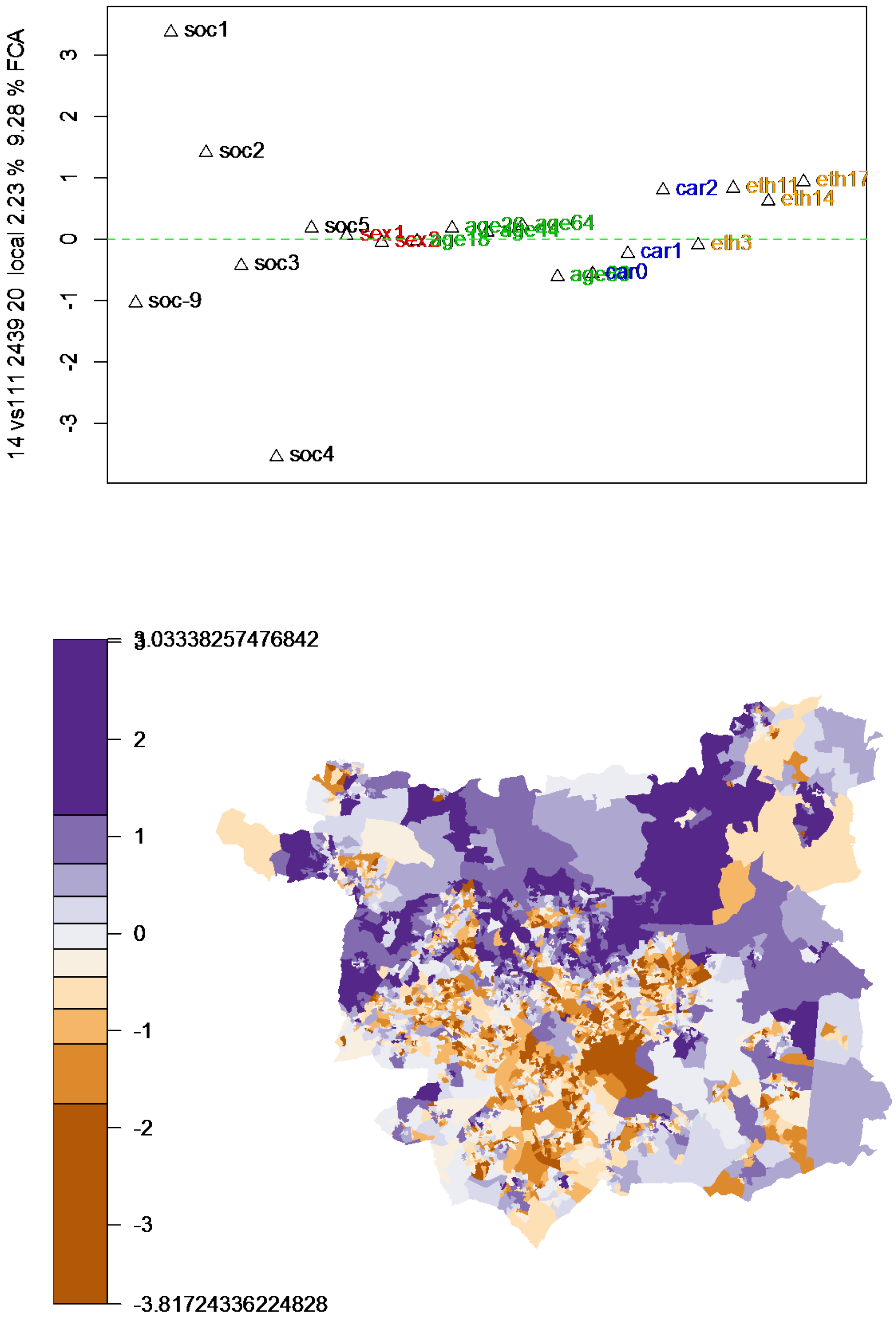

Figure 19: Principal Tensor n°32 of the FCAk

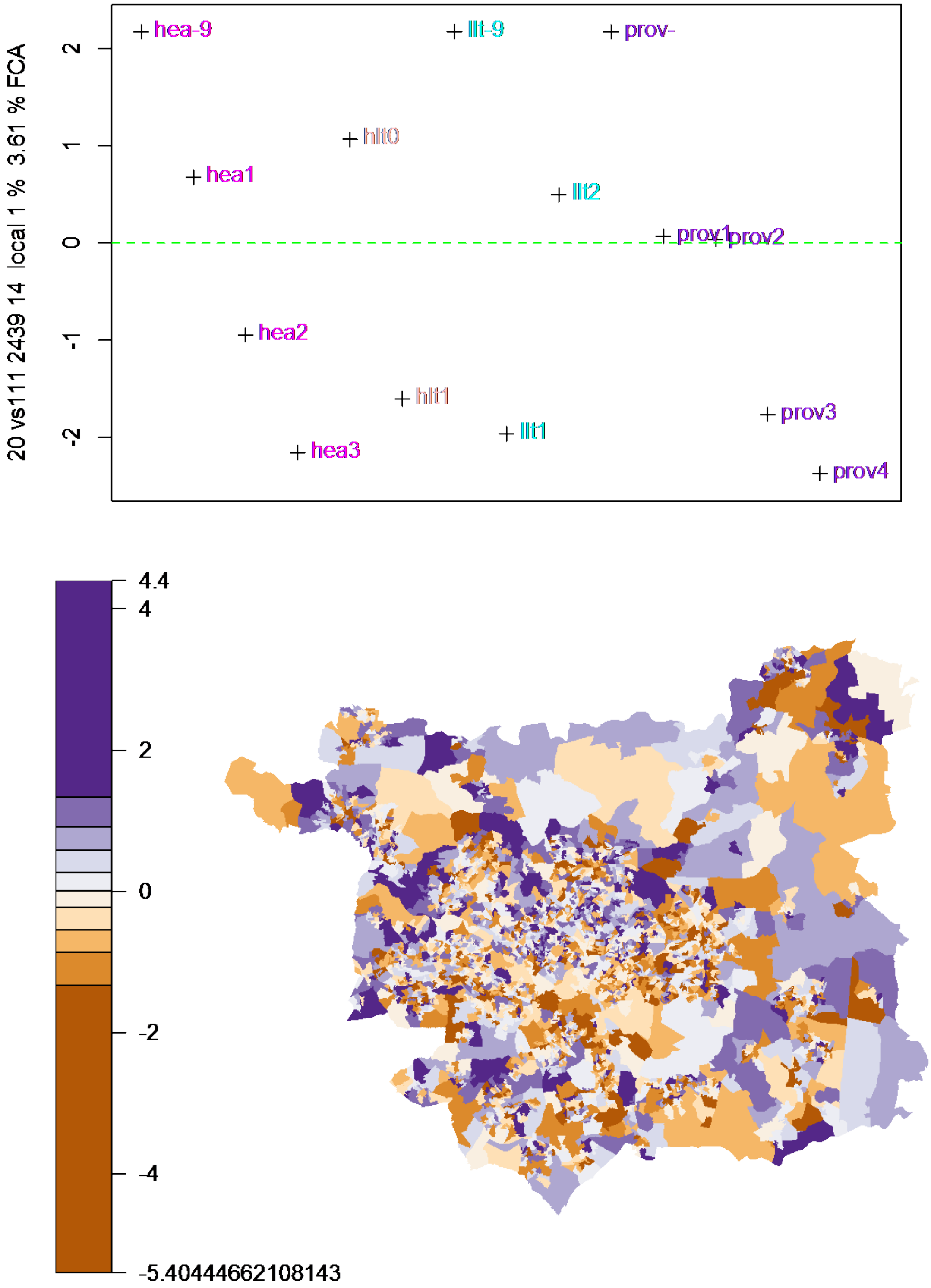

Figure 20: Principal Tensor n°17 of the FCAk

Within the same part of the decomposition (a two-way analysis once contracted by the marginal health domain), the tensor n°32 (Figure 19) expressing 9% of lack of independence shows the social grade spatial effect, opposing manual less skilled workers (soc4) in the city center and the west of the district to the non- manual workers (soc1 and sco2) in the north and eastern parts. This analysis can be done with time as fourth mode to be able to capture some evolution of these spatial associations within and between the domains.

5 Discussion and conclusion

This paper has applied an extended range of methods related to correspondence analysis to a (micro-) simulation allowing rich representations of the socio-demographic structure of a city, as it is now and as it may be expected to evolve into the future. It is typical of simulations of this type that substantial volumes of data are generated – in this case, an array of 2439 areas with 39 attributes and 7 time periods gives more than half a million counts for the analysis. Detection of trends and patterns in a dataset of this type is a non-trivial challenge, but can potentially be valuable in pointing to substantive change (which may also have policy relevance) and also be suggestive of the function, and in some case perhaps the malfunction, of the underlying models.

Starting with a simple correspondence analysis, we found that a representative pair of characteristics – social class and health – are far from spatially independent. In addition to a strong association between professional occupations and good health, and between the unemployed and socially inactive with poor health, a notable tendency was detected in the failure of the higher social grades to report health status. This was extended to a multiple correspondence analysis in which a group of eight attributes from the Moses microsimulation model were combined simultaneously. This appeared to show associations not just between health and social status but also indicators such as lack of car ownership, membership of ethnic minority groups and stage in the family lifecycle. Altogether this model appears to provide a multidimensional assessment of neighbourhood quality, or perhaps even 'deprivation', across the city of Leeds. Map-based representations of this indicator (Figure 5) show a typical in-out pattern of location quality. By comparing an MCA model at a finer spatial scale (LSOA) against a model of independence at a higher spatial scale (wards) it was also possible to represent homogeneity against multiple criteria. This analysis is suggestive that the more spatially aggregated simulations of the earlier Moses implementations (*e.g.*, Birkin et al, 2006) will benefit significantly from further disaggregation to a neighbourhood scale.

The power of a multi-way analysis was seen through the incorporation of both spatial scale and temporal evolution alongside spatial variations in multiple attributes. The patterns of spatial variation in 'deprivation' and social grade appeared to be robust in relation to both scale and time, and indeed it was possible to suggest a composite map of neighbourhood quality across the entire horizon of the simulation (Figure 15). However the analysis also seemed to show a progressive reduction in the differentiation between areas during the course of the simulation, and this may be indicative of a failure to fully incorporate local migration preferences in the version of the MSM that has been considered here. Thus, for example, Jordan et al. (2011) have suggested the addition of segregation rules in the style of Schelling to improve the movement model here. This trend was seen more clearly in the inner suburbs of the city than the surrounding towns (see Figure 12: inner suburb wards number are 4, 33, 2, 5, 14, 9, 30, 18, 27, 29 and 11) but this is perhaps not unreasonable. Recent trends suggest relatively low migration rates within the city itself, while the process of demographic ageing and the associated social adjustments (*e.g.,* reductions in household size, car ownership and so on) will likely be more pronounced in the outlying towns and villages.

In a final set of investigations, the health variable was again isolated for special attention within the multi-way framework. This appeared to present some fresh perspectives on the question of health geographies – in effect a synoptic view of 'healthy spaces' abstracted from both scale and time (Figure 18). Such analysis could be thought-provoking and instructive when considering indicators of provision of health care – both current and planned – and might also be usefully extended to other domains such as housing or education.

6 References

Ballas, D., & Clarke, G. (2000). GIS and microsimulation for local labour market policy analysis. *Computers, Environment and Urban Systems, 24*, 305 –330.

Ballas, D., Clarke, G., Dorling, D., Eyre, H., Thomas, B., and Rossiter, D. (2005) 'SimBritain: A Spatial Microsimulation Approach to Population Dynamics. *Population, Space and Place*, 11, 13-34.

Birkin, M.H. Clarke, M. (1987) Comprehensive models and efficient accounting frameworks for urban and regional systems. In Griffith, D., and Haining, R. (Eds) *Transformations through space and time*, Martinus Nijhoff, The Hague, 169-195.

Birkin, M., Turner, A., and Wu, B. (2006) A synthetic demographic model of the UK population: Methods, progress and problems. *In:* Regional Science Association International British and Irish Section, 36th Annual Conference, The Royal Hotel, St Helier, Jersey, Channel Islands.

Birkin, M.H., Townend, P., Turner, A., Wu, B. and Xu, J. (2009) MoSeS: A Grid-enabled spatial decision support system. *Social Science Computing Review*, 27, 4, 493-508

Birkin, M.H. and Clarke, M. (2011) Spatial microsimulation models: A review and glimpse into the future. *In:* J. Stillwell, M. Clarke (Eds.), Population dynamics and projection methods. *Understanding population trends and processes*, Vol. 4 Springer, Dordrecht, pp. 193–208

Dray, S. and Dufour, A.B. (2007) The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, vol. 22, no. 4, p. 120

Escofier, B. (1984) Analyse factorielle en référence à un modèle. Application à l'analyse de tableaux d'échanges. *Revue de Statistique Appliquée*, vol. 32, no. 4, pp. 25–36. Retrieved July 24, 2012, from http://www.numdam.org/item?id=RSA_1984__32_4_25_0

Gatrell, A.C., Popay, J. and Thomas, C. (2004) Mapping the determinants of health inequalities in social space: can Bourdieu help us? *Health & Place*, vol. 10, no. 3, pp. 245–257.

Greenacre, M.J. (2007) *Correspondence Analysis in Practice*, CRC Press.

Hermes, K. and Poulsen, M. (2012) A review of current methods to generate synthetic spatial microdata using reweighting and future directions. *Computers, Environment and Urban Systems*, vol. 36, no. 4, pp. 281–290.

Jordan, R., Birkin, M.H. and Evans, A. (2011) Agent-Based Simulation Modelling of Housing Choice and Urban Regeneration Policy. *In*: Bosse, T., Geller, A., Jonker, C.M. (Eds.), *Multi-Agent-Based Simulation XI*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 152–166.

Lebart, L., Morineau, A. and Warwick, K.M. (1984) *Multivariate descriptive statistical analysis: correspondence analysis and related techniques for large matrices*. Wiley.

Leibovici, D.G. and El Maâche, H. (1997) Une Décomposition en Valeurs Singulières d'un Elément d'un Produit Tensoriel de k Espaces de Hilbert Séparables. *Compte Rendus de l'Académie des Sciences* I, 325(7), 779–782.

Leibovici, D.G. (2010) Spatio-Temporal Multiway Data Decomposition Using Principal Tensor Analysis on k-Modes: The R Package PTAk. *Journal of Statistical Software*, 34(10), pp.1–34. Available at: http://www.jstatsoft.org/v34/i10.

Leibovici, D.G. and Jackson, M. (2011) Multi-scale Integration for Spatio-Temporal Ecoregioning Delineation. *International Journal of Image and Data Fusion, 2(2): 105-119*

Leibovici, D.G. and Birkin, M.H. (2013) Geocomputational Perspectives for Entropic Variations of Urban Dynamics. *Geographical Analysis (submitted)*

Le Roux, B., Rouanet, H., Savage, M. and Warde, A. (2008) Class and Cultural Division in the UK. *Sociology*, vol. 42, no. 6, pp. 1049–1071.

Nenadic, O. and Greenacre, M .(2007) Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca Package. *Journal of Statistical Software*, vol. 20, no. 3, pp. 1–13. Available at: http://www.jstatsoft.org/v20/i03

Murtagh, F. (2005) *Correspondence Analysis And Data Coding With Java And R*. CRC Press.

Openshaw, S. (1983) *The modifiable areal unit problem*. *Concepts and Techniques in Modern Geography (CATMOD)*, 38, Geo Books Norwich,UK.

Orcutt, G. (1957). A new type of socio-economic system. *Review of Economics & Statistics*, 58, 773–797.

Procter, K., Clarke, G., Ransley, J., and Cade, J. (2008). Micro-level analysis of childhood obesity, diet, physical activity, residential socio-economic and social capital variables: Where are the obesogenic environments in Leeds? *Area, 40*(3), 323 –340.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http: //www.R-project.org/.

Savage, M. (2010) (Focus Article) The Politics of Elective Belonging. *Housing, Theory and Society*, vol. 27, no. 2, pp. 115–161 (with discussions)

Smith, D. M., Clarke, G. P., Ransley, J., and Cade, J. (2006). Food access & health: A microsimulation framework for analysis. *Studies in Regional Science*, 35(4), 909–927.

Statistics Canada. (2009). *ModGen developers guide*. Online at http://www.statcan.gc.ca/microsimulation/modgen/modgen-eng.htm

Tanton, R., McNamara, J., Harding, A., and Morrison, T. (2009). Small area poverty estimates for Australia's Eastern Seaboard in 2006. In A. Zaidi, A. Harding, & P. Williamson (Eds.), *New frontiers in microsimulation modelling* (pp. 79 –96). Farnham: Ashgate.

Nakaya, T., Fotheringham, A. S., Clarke, G., & Ballas, D. (2007). Retail modelling combining meso & micro approaches. *Journal of Geographical Systems*, 9, 345–369.

Townend, P., Xu, J., Birkin, M.H., Turner, A. and Wu, B. (2009) Modelling and Simulation for e-Social Science. *Philosophical Transactions of the Royal Society* A, 367, 2781-2792

Van Imhoff, E. and Post, W. 1998 Microsimulation methods for population projection. *Popul. Engl. Select.* **10**, 97–138.

Wu, B. M., and Birkin, M.H. (2012) Agent-Based Extensions to a Spatial Microsimulation Model of Demogaphic Change. *In: A.J. Heppenstall et al. (eds) Agent-Based Models of Geographical Systems*, Springer Sciences Business Media ,347-360

7 Appendix

The 10 variables in the article and used in the MoSes data simulation, are description in the codebook of the 2001 Individual Licensed SAR Version 2.5, www.ccsr.ac.uk/sars.: "The 2001 Individual Licensed SAR (IL-SAR) is a 3 per cent sample and contains over 1.75 million records. It contains a full range of census topics on individuals and summary information about households and the new information collected on qualifications, caring and religion. Geographical information is given down to Government Office Region."

For each variable is presented: - the labelling of categories, the variable name in the SAR with its definition - the values, with possible recoding used in this article in a table extracted from the SAR document with percentages given in the SAR (as is or a sum for some categories for recoded variables).

age<code> *age0 Age of respondent*

group	code	Percentage
<=18	18	21.34%
18-26	26	17%
26-44	44	23%
44-64	64	23.59 %
>64	80	15.07%

sex<value> *sex Sex of the respondent*

Value	Label	Percentage
1	Male	48.65%
2	Female	51.35%

soc<value> *hrsocgrd social grade of the household reference person*

Value	*Label*	*Percentage*
-9	Not applicable(not in a hhd, student living away, hrp not 16)	8.20%
1	A/B Professional Middle managers	22.16%
2	C1 All other non-manual workers	24.90%
3	C2 All skilled manual workers	16.21%
4	D All semi-skilled and unskilled manual workers	18.53%
5	E On benefit/unemployed	10.00%

car<code> *car0 Number of cars owned or available for use in the household*

Value	code	Label	Percentage
-9	0	Not applicable (not in a household)	1.79%
0	0	No car	19.44%
1	1	1 car	41.30%
2	2	2 cars	29.14%
3	2	3 cars or more	8.33%

eth<code> *ethw Ethnic group for England and Wales reclassified in 5 groups*

Value	code	Label	Percentage
-9	-9	Not applicable (Scot/NI)	12.65%
1	3	British	76.47%
2	3	Irish	1.06%
3	3	Other White	2.24%
4	14	White and Black Caribbean	0.39%
5	17	White and Black African	0.13%
6	11	White and Asian	0.31%
7	17	Other Mixed	0.26%
8	11	Indian	1.73%
9	11	Pakistani	1.20%
10	11	Bangladeshi	0.47%
11	11	Other Asian	0.40%
12	14	Black Caribbean	0.94%
13	14	Black African	0.81%
14	14	Other Black	0.16%
15	17	Chinese	0.38%
16	17	Other Ethnic Group	0.37%

hea<value> *health Self-assessment for general health in the last 12 months*

Value	Label	Percentage

-9	Not applicable (student living away)	0.97%
1	Good	67.84%
2	Fairly good	21.95%
3	Not good	9.24%

lti<value> *llti* *Limited long term illness*

Value	Label	Percentage
-9	Not applicable (student living away)	0.97%
1	Yes	18.27%
2	No	80.76%

hlt<code> *hhlthind* *Household health and disability indicator*

Value	code	Label	Percentage
-9	0	Not applicable (not in a household)	1.79%
0	0	Noone in hhd lt-ill or in poor health	63.89%
1	1	Hhd member lt-ill or in poor health	34.32%

pro<value> *provcare* *Number of hours (unpaid) care provided per week*

Value	Label	Percentage
-9	Not applicable (student living away)	0.97%
1	Provides no care	89.12%
2	Proves 1-19 hours care	6.68%
3	Provides 20-49 hours care	1.12%
4	Provides 50 or more hours care	2.11%

trw<code> *tranwrk0–* transport to work

Value	code	Label	Percentage
-9	-9	Not applicable (not aged 16 to 74 or not working or student)	53.81%
1	1	Work mainly from home	3.77%
2	3	Underground metro light rail(EW&S) or tram(E&W) or tube(S)	1.20%
3	3	Train	1.80%
4	3	Bus minibus or coach	3.90%
5	6	Motor cycle scooter or moped	0.46%
6	6	Driving a car or van	24.73%
7	6	Passenger in a car or van	3.26%
8	6	Taxi or minicab	0.27%
9	9	Bicycle	1.18%
10	9	On foot	5.30%
11	9	Other	0.25%
12	6	Car or van pool for NI only	0.07%