Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Home Affairs FDHA
Federal Statistical Office FSO

Work for the Swiss Federal Statistical Office

Small Area Estimation in the Structural Survey

Report 1.1

EXPERIMENTAL STATISTICS

Neuchâtel, 2018

Published by:	Federal Statistical Office (FSO)	**Layout concept:**	Section DIAM
Information:	info.pop@bfs.admin.ch, tel. +41 58 467 25 25	**Downloads:**	www.statistics.ch
Editors:	Section METH, FSO	**Copyright:**	FSO, Neuchâtel 2018 Reproduction with mention of source authorised (except for commercial purposes)
Contents:	Ewa Strzalkowska and Isabel Molina Department of Statistics, Universidad Carlos III de Madrid		
Topic:	00 Statistical Basis and Overviews		
Original text:	English		

ESPACE DE L'EUROPE 10 CH-2010 NEUCHÂTEL WWW.STATISTICS.CH

Small Area Estimation in the Structural Survey: Phase I Work for the Swiss Federal Statistical Office

Ewa Strzalkowska and Isabel Molina
Department of Statistics, Universidad Carlos III de Madrid

March 3, 2014

1 Introduction

Phase I of this project consists of analyzing the performance of different small area estimators of the proportions of active people in the Swiss districts using the data set STATPOP provided by the Swiss Federal Statistical Office. Concretely, we want to reproduce the simulation experiments carried out previously for the Spanish Labor Force Survey data using now the provided Swiss data, with the purpose of answering the following questions:

(a) Can we find a regression model that explains adequately the activity in Swiss districts?

(b) Are synthetic estimators that do not consider random district effects good enough for estimation of the proportions of active people in the districts, or do we need to consider models with random district effects (mixed models)?

(c) Are plug-in estimators visibly biased? In such a case, we would need to obtain empirical best predictors by Monte Carlo approximation.

(d) Are linear models (LMs) good enough for estimation of the proportions of active people, or do we need to consider exclusively generalized linear models (GLMs)? An adequate GLM for a dummy dependent variable is a Binomial model with logit link (logistic GLM).

(e) Are the provided covariates good enough when using an area level model such as the Fay-Herriot model?

The analysis will be based on several simulation studies in which the population, and hence the true values of the district proportions of active people, are known. A subset of the data set STATPOP will play the role of "population". Concretely, we consider the records of STATPOP that are contained in the Structural Survey, because the variable of interest (active/not active) is available only for this survey. Since the survey data set is regarded as the population, direct estimates act as true values. Comparisons of the

different estimates with those "true values" will be fair only if the sample sizes of the districts are large enough for direct estimates to be reliable. For this reason, only the districts with a minimum sample size in the Structural survey were considered; concretely, we selected the districts in which direct estimators have coefficients of variation below 10%. The final subset that plays the role of the population in simulation studies is called here STRPOP.

Once the desired data set STRPOP is selected in Section 3, in Section 4 we search for the best possible GLM model that explains the true district proportions of active people in terms of the available auxiliary variables for those data. A previous process of selection of the suitable auxiliary variables for the model is carried out. The final model contains fixed district effects because the districts retained in the sample have sufficient sample sizes for estimating these district effects. The goodness of this model is analyzed by generating a population from the model and comparing the district proportions for this "theoretical" population, which follows exactly the model, with the "true" proportions from STRPOP. This theoretical population is called hereafter STRPOP*.

In a first experiment described in Section 5, we draw a single sample from each population, the original one (STRPOP) and the "theoretical" population (STRPOP*). The sample units are the same in both samples. The sample is composed of D independent simple random samples drawn from the D districts that were considered in STRPOP. In this way we are making sure that the sampling design has no influence at all in the estimation process and the performance of estimators can be compared in absence of possible distortions caused by the sampling design. The district sample sizes are taken such that the smallest size is approximately equal to the smallest sample size in the original Structural Survey. This allows us to study the behavior of the estimators for districts with sample sizes as small as those smallest districts the original Structural Survey. Using the resulting sample from each population, we calculate the same estimators considered in the previous simulation study using Spanish data. The performance of the estimators for this single sample is compared with true values under each population, STRPOP and STRPOP*.

A second simulation experiment (Section 6) is performed under the "model-based" setup, in which the performance of estimators is analyzed by averaging over all the possible populations generated from the considered GLM model. This entails generating a large number of populations from the fitted model, taking a sample from each population (all samples have the same units), calculating all the estimators from each sample data and calculating performance measures by averaging over the simulated populations. Note that in this simulation setup, true proportions are varying over simulations.

A third simulation experiment (Section 7) is performed under the "design-based" setup but assuming that the population follows a GLM model to avoid potential model failure distortions. In "design-based" simulation experiments, the performance of estimators is analyzed by averaging over all the possible samples drawn from a fixed population. In this case, we consider the STRPOP* population generated from the GLM model with fixed district effects. Thus, we take the population STRPOP*, let it fixed and draw a large number of samples from it. All the estimators are calculated from each sample data and performance measures are calculated by averaging over the different samples. In this

case, true proportions are fixed and only estimates are varying over simulations.

Finally, the last simulation experiment is performed by drawing samples from the original data set STRPOP, which does not strictly follow the GLM model. This experiment allows us to analyze the effect of potential model failures on the final estimators under a design-based setup.

2 Data description

As already mentioned, in Phase I of this project we consider only the registers of the STATPOP data set that are available in the Structural Survey. The sample size of the Structural Survey is $286,015$ out of the total size of the STATPOP data set $N = 6,662,333$. The number of districts is 147. The sample and population sizes of the districts in the Structural Survey are listed in Table1.

Table 1: Sample and population sizes of Swiss districts

Popn. size (N_d)	nr. of districts	Average sample size (n_d ave.)
1839-3000	2	76
3,000-5,000	6	151.6
5,001-10,000	11	331.1
10,001-50,000	88	1,185.8
50,001-200,000	37	3,685.7
200,000- 352,950	3	13,530

All variables included in the STATPOP data are listed in Appendix 2. A new set of variables has been constructed by recoding, putting together categories or combining several variables from the STATPOP data. Specifically:

- The variables “populationtype” and “TypeofHousehold” were excluded from the study since all observations are equal to 1.
- The original variables ”nationalityid” and ”Residencepermit’ were combined into one variable called ”nationality” (with categories Swiss/NonSwiss).
- In the original variable “maritalStatus”, three of the categories, concretely “unmarried”, “in a registered partnership” and “partnership dissolved” contained just few observations. These three categories were put together with “single”, “married” and “divorced” respectively. The new variable is named “civil status”.
- The 12 variables “January“–“December” were aggregated into a single one called “OASI in one year”. This variable has three categories: 1 - no contribution to OASI during the whole 2011, 2 - the person contributed to OASI only part of the year (between 1 and 11 months) and 3 - the person contributed to OASI the whole year 2011.

- The variable "Income" is registered only for persons who contributed to OASI, hence it needed to be categorized. We constructed a new variable "Income" with 4 categories: 1 - unknown income, 2 - income less than the mean of the observed incomes (up to 24,000 CHF), 3 - income between 24,000 CHF and 60,000, and 4 - income above 60,000.

- The "HouseholdSize" was reduced only to 4 categories: 1 - households of size one, 2 - households of size two, 3 - households of size between 3 and 5, 4 - households of size between 6 and 10 and 5 - larger households/communes (more than 10).

- The original variable "secondaryResidenceId1" was transformed into "Secondary Residence", which has only two categories (yes and no).

- The variable "age" was categorized as described in Table 2. It is important to mention that using the original variable "age" as a continuous variable was giving worse results in our study.

- Finally, note that we have the following relation between different types of Swiss regions: Municipalities$\subset$ Districts$\subset$ Strata$\subset$Cantons$\subset$NUTS2. In our study the districts are the areas of interest. The Strata will be also considered in the models so as to include the design information. The other variables were not considered in our simulation studies.

Table 2 lists the final set of variables that will be considered in our analysis.

Table 2: Considered variables

active	1=active 0=inactive
district	there are 147 of them
Strata	there are 27 of them (include 1 or more districts, no overlapping)
age group	$[15,20)$, $[20,60)$, $[60,65)$, ≥ 65
gender	male (1) / female (0)
nationality	Not swiss (1) / Swiss (2)
civil status	1 = single, unmarried, 2 = married, in a registered partnership, 3 = widow/er, 4 = divorced, partnership dissolved
OASI in one year	1=no contribution, 2= contr. part of year, 3= contr. full year
In OASI	yes/no
Income	unknown (In OASI = no), $(0,24000]$, $(24000,60000]$, > 60000
Household Size	1, 2, [3,5], [6,10],>10
secondary residence	no (1) / yes (2)
weight	Design weight of the unit

3 Population data for simulations

let Y_{di} denote the variable taking value 1 if i-th person in d-th district is active and 0 otherwise. The true proportion of active people in district d is defined as

$$P_d = \frac{1}{N_d} \sum_{i=1}^{N_d} Y_{di},$$

where N_d is the population size of district d. Now let s_d be the subsample from district d in the Structural Survey, of size $n_d < N_d$, and w_{di} be the sampling weight of i-th person in d-th district in that Survey. A direct estimator of P_d is the Horvitz-Thompson estimator given by

$$\hat{P}_d = \frac{1}{N_d} \sum_{i \in s_d} w_{id} Y_{di}.$$

Using the approximation $\pi_{di,dj} \approx \pi_{di}\pi_{dj}$ for second order inclusion probabilities in terms of first order inclusion probabilities, which is exact for Poisson sampling, an estimator of the design-variance of $\hat{P}_d$ is given by

$$\hat{V}_\pi(\hat{P}_d) = \frac{1}{N_d} \sum_{i \in s_d} w_{di}(w_{di} - 1) Y_{di}^2.$$

For each district in the Structural Survey, we calculate direct estimates $\hat{P}_d$ of the proportions of active people and their estimated coefficients of variation (CVs), given by $\hat{CV}_\pi(\hat{P}_d) = 100 \times \hat{V}_\pi(\hat{P}_d)/\hat{P}_d$. There happen to be 6 out of the 147 districts with CV greater than 10%, see Table 3 for a description of these districts. We include in the data set STRPOP the remaining $D = 141$ districts, whose direct estimators have CVs smaller than 10%. This data set is treated hereafter as the population. The "population" size in STRPOP is $N = 285,377$. The population sizes of the districts vary from $\min(N_d) = 162$ for the smallest to $\max(N_d) = 20,786$ for the largest district. Figure 1 shows a boxplot and a histogram of these population sizes. These plots show that the are few very large districts as compared with the other ones.

Table 3: Removed districts

District id	502	1401	1404	1406	1822	2304
Sample size	62	106	90	122	120	138
CV	16.02	11.50	12.90	10.49	11.86	11.13

4 Population model

The goal of this section is to find a model that explains adequately the activity in Swiss districts with the available data. This model will be used to generate a theoretical population, called STRPOP*, that follows exactly the model and should be similar to the STRPOP

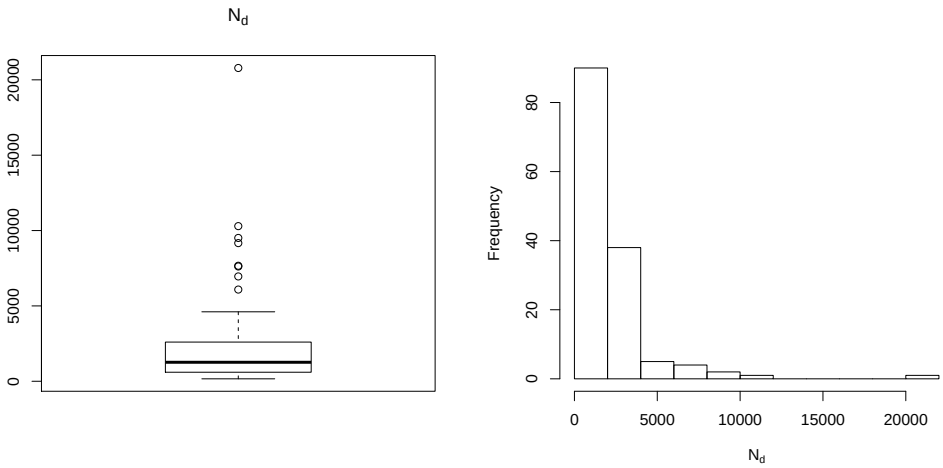

Figure 1: Boxplot and histogram of district population sizes in STRPOP.

data if the model is approximately correct. Comparison of the true proportions of active people in the two populations, STRPOP and STRPOP*, will give us an indication about the validity of the model.

Recall that Y_{di} takes value 1 if i-th person in d-th area is active and 0 otherwise. A suitable model for a binary response variable is a generalized linear model (GLM) with Binomial distribution and logit link. Let p_{di} be the true probability of being active and $\mathbf{x}_{di}$ a vector containing the values of the explanatory variables for that same unit. The following GLM with fixed district effects was fitted to the STRPOP data:

$$Y_{di} \sim Bern(p_{di}), \tag{1}$$

$$\log\{p_{di}/(1+p_{di})\} = \mathbf{x}'_{di}\beta + \alpha_d,\ i = 1,\ldots,N_d,\ d = 1,\ldots,D, \tag{2}$$

where α_d is the fixed effect of district d and β is the vector of regression coefficients of the auxiliary variables. We consider fixed district effects in order to reproduce in a nonparametric way the real district effects in the data and also because the samples sizes of the selected districts in STRPOP are not so small and therefore the estimated fixed effects for these districts will be minimally reliable.

As goodness-of-fit criteria to select the adequate explanatory variables, we consider: log-likelihood, AIC, BIC and correct classification rates of actives and non-actives at the national level. Note that the contribution to OASI is a potentially important explanatory variable for the activity. See in Table 2 that there are three variables, “OASI in one year”, “In OASI” and “Income”, which include the same information concerning whether the unit contributed to OASI or not. Thus, only one of them should be included in the model. To select which one explains best the activity, we fitted the GLM with each of the three variables as explanatory variables. Results were:

- Including “OASI in one year” gives us AIC: 155,375.
- Including “In OASI” gives us AIC: 160,963.

- Including “Income” gives us AIC: 152,145.

Since the variable “Income” gives a better fit according to the AIC goodness-of-fit criteria, we consider “Income” as explanatory variable.

The sampling weight should not be significant once the model accounts for the only design variable, the sampling strata. Significancy of the sampling weight would suggest that a relevant design variable is missing in the model in an adequate form. A plot of the distribution of the sampling weights across strata in Figure 2 reveals two groups of Strata with a clearly different median weight. Thus, including fixed effects for these two groups of strata should explain sufficiently well the information provided by the sampling weights. Districts, however, are nested within Strata and the model already includes fixed effects for the districts. District effects already incorporate the strata grouping effect and hence, including in the model the district effects is good enough for explaining the sampling weights in the population model. In the models that will be fitted for the sample in later sections, fixed district effects are not included because district sample sizes are too small, and consequently the mentioned grouping variable of the strata should be included.

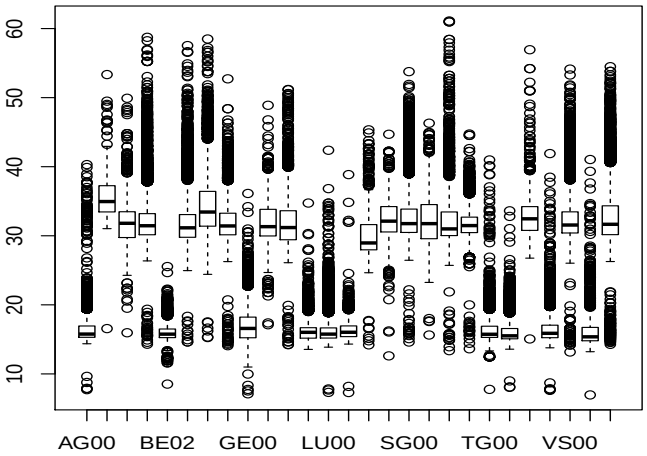

Figure 2: Sampling weights for Strata.

The final GLM model (1)-(2) that was fitted to the STRPOP data included all the variables listed in Table 4, except for `Strata1` that is considered only for the sample models. The interactions `gender*age group` and `gender*civil status` were also included in the model. All considered explanatory variables and interactions are strongly significant, except for at most one of the categories of some of the variables or interactions. Using this model, the failure classification rate for the outcome (Active/Non active) at the national level is 11.50%.

Table 4: Variables (final)

active	1=active 0=inactive
district	there are 147 of them
Strata1	0=Strata with large sampling weight median, 1=otherwise
age group	$[15,20)$, $[20,60)$, $[60,65)$, ≥ 65
gender	male (1) / female (2)
nationality	Not swiss (1) / Swiss (2)
civil status	1 = single, unmarried, 2 = married, in a registered partnership, 3 = widow/er, 4 = divorced, partnership dissolved
Income	unknown (In OASI = no), $(0,24000]$, $(24000,60000]$, > 60000
Household Size	1, 2, [3,5], [6,10] >10
secondary residence	no (1) / yes (2)

5 Simulation under fixed population and sample

A theoretical population, called STRPOP*, was generated from the fitted model (1)-(2). The district population sizes were the same as in the original population STRPOP. This was done simply by generating Bernoulli random variables with true probabilities equal to the resulting fitted values from the GLM fit to the original data in STRPOP. The new theoretical population STRPOP* follows exactly the GLM model. Then estimators will be compared under a population that exactly follows this model, which helps to avoid confounding estimation errors from model failure errors.

We have two populations, the original data STRPOP and the theoretical population STRPOP*. Each population has a set of true proportions of active people in the districts. Figure 3 plots the true proportions from STRPOP against those in the theoretical population STRPOP*. We can see that the points lie around the line, which indicates some similarity between the two sets of proportions, although the similarity is weaker than in the previous experiment for Spanish data. Although the considered GLM model seems to have a slightly weaker predictive power than that one considered for the Spanish data, we believe that this GLM model seems to explain acceptably well the activity in the Swiss districts, which answers positively question (a). However, we remark that the activity in Swiss districts seems to be explained to some extent by factors other than those considered in the model.

A sample was drawn from each of the two populations. The sample was drawn by stratified random sampling with districts as strata and simple random sample without replacement within each district. This will also allow to study estimation errors that are not due to the fact of having different sampling weights within the same district. Thus, we have two parallel populations with their corresponding sample. The sample size for the smallest district is taken to be 60, which is approximately the smallest sample size in the Structural Survey. For the second largest district, the sample size was incremented by 5 and so on and so forth until arriving to the sample size 760 for the largest district. The sample indices and the auxiliary variables are the same for the two populations. For

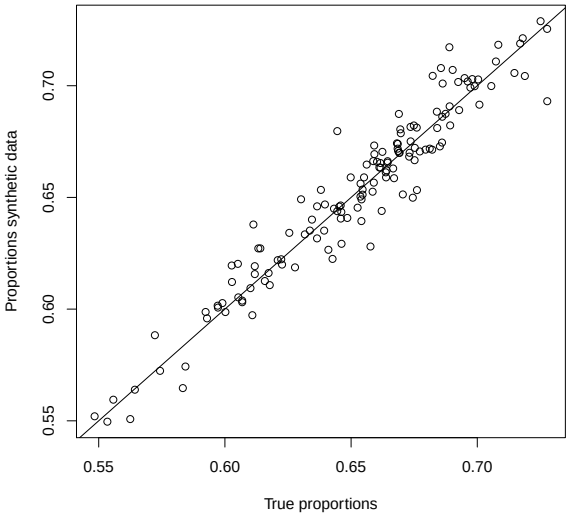

Figure 3: True area proportions of active people in the original data set STRPOP and in the theoretical population data STRPOP*.

each sample, we calculated all the estimators listed in Table 5 and described in Appendix 1. Predictive estimators, which are based on predicting also sample values along with the non-sample ones, are not included in this study because the alternative estimators that predict only the non-sample data and preserve the sample data are theoretically closer to the best estimators. With the available data, linking the structural survey with the register from which auxiliary variables are obtained is not a problem and therefore predictive estimators are not necessary. All the models included in Table 5 were fitted using the same set of auxiliary variables. These auxiliary variables are those considered in the model for the population, but including the grouping strata variable `Strata1` and removing the district fixed effects because sample sizes are too small for several districts and therefore fixed district effects cannot be estimated efficiently for those districts; doing it would lead to inefficient estimators of the target district proportions. GLM and LM models do not include district effects at all and they are called "synthetic", whereas GLMM and LMM include random district effects.

Hereafter we compare the estimators listed in Table 5 for the given populations and corresponding samples and we will try to answer the questions posed in the Introduction. In later sections, estimators will be compared by averaging their performance across different populations generated from the same population model (model-based setup), across different samples drawn from a fixed population that follows the population model (design-based setup, knowing the data generation process), and across different samples drawn from the original population (design-based setup, not knowing the data generating process).

Thus, in the remainder of this section we describe the results for fixed population and

Table 5: Estimators considered in the simulation study.

Point estimator	Area/Unit level	Model
Direct (HT)	No model	No model
GLM (synthetic) (plug-in)	Unit level	Generalized Linear Model (Binomial, logit link)
GLMM (plug-in)	Unit level	Generalized Linear Mixed Model (Binomial, logit link)
LM (synthetic)	Unit level	Linear Model (no area effects)
LMM	Unit level	Linear Mixed Model (BHF model)
FH	Area level	Linear Mixed Model (FH model)

sample. Figure 4 plots direct estimates against true values, for theoretical data STRPOP* and for true data STRPOP. See that points form a band around the line of equality in the two populations but they are not very close to the line for this sample. Direct estimates seem to have a wider range of variation than true values, which suggests that they are not very efficient, especially for the districts with smallest sample sizes.

Concerning the estimators based on models, first we take a look at the goodness-of-fit of each model in terms of AIC. We need to remark that the AIC is comparable only for models belonging to the same family. The GLM is a particular case of the GLMM with random effects variance equal to zero and therefore these two models are comparable. However, according to Table 6, the AIC does not really discern among one of them for the true data. For the theoretical data, Table 7 indicates that the GLMM model fits better according to the AIC. LM an LMM are the linear analogues of GLM and GLMM. In this case, LM gets a better AIC for the two populations. We must mention that the real aim of the models in this study is in predicting rather than explaining the relationships between the activity and the auxiliary variables. Thus, goodness-of-fit is used here only as an indication, but comparison of estimates with true values is much more relevant in this case.

Table 6: Akaike Information Criteria (STRPOP)

GLM	GLMM	LM	LMM
31156.00	31145.82	20179.00	20376.52

Table 7: Akaike Information Criteria (STRPOP*)

GLM	GLMM	LM	LMM
30712.00	30681.09	19030.00	19208.34

As next, we analyse the plug-in estimators based on the synthetic GLM model without

district effects. In Figure 5 we plot these estimates against true values for the two populations, STRPOP* and STRPOP. Points are much closer to the line than in the previous plot. Thus, for this sample, estimates based on the synthetic GLM model are definitely better than direct estimates. Figure 6 shows the analogous plot for the plug-in estimates based on the GLMM, which includes random district effects. These plots look somewhat similar to the previous ones for the synthetic GLM model and no relevant bias appears for any of the populations, which answers negatively question (c). The corresponding plots for FH estimates are shown in Figure 7. These plots also look good but we will see later that FH estimates are not as efficient as other considered estimators when averaging across populations or across samples.

Now to answer question (d) on whether linear models are good enough or we have to resort to generalized linear models, we compare the synthetic LM with the plug-in estimates based on the synthetic GLM. In this case we calculated the percent relative error (RE) of each type of estimate. The percent RE of an estimate $\hat{P}_d$ of the true proportion P_d is calculated as

$$RE(\hat{P}_d) = 100\frac{\hat{P}_d - P_d}{P_d}.$$

Figure 8 plots the REs of estimators based on the synthetic GLM against those obtained from the synthetic LM. Observe that relative errors of estimators based on the GLM, which is specific for binary data, are really very similar to those based on the LM. This conclusion remains true for the models that incorporate random district effects, see Figure 9. The reason is that the logistic function is approximately linear for probabilities that are in the interval (0.2,0.8). Since the true district proportions of active people are not far from 0.5, the GLM with logit link is approximately equal to the linear model when estimating these proportions.

Additionally, Figure 10 compares the percent REs of the estimates based on the GLMM against those of the FH estimates. See that most points are on the left side of the plots, which means that for this sample FH estimates have larger RE than the plug-in estimates based on the GLMM for most districts.

The REs of the estimates based on the GLM, GLMM, LM and LMM, along with FH estimates, are plotted in Figure 11 for each district, with districts sorted in the increasing order of their sample size. See that direct estimates have larger relative errors for most districts, followed by FH estimates. The other four estimates get very similar results, but models with random district effects (LMM and GLMM) tend to decrease slightly the largest errors. Thus, the answer to question (b) on whether synthetic models are good enough or we should consider models with random district effects, is that it is more recommendable to use models with random effects. In order to see the differences among estimators more clearly, Figure 12 plots the REs of only direct, FH and plug-in estimates based on the GLMM with random district effects. Although FH estimates improve the direct estimates in practically all districts, the reductions in RE are clearly more striking for the estimates based on the GLMM. All these conclusions will be corroborated when averaging over a large number of possible populations or samples in the next sections.

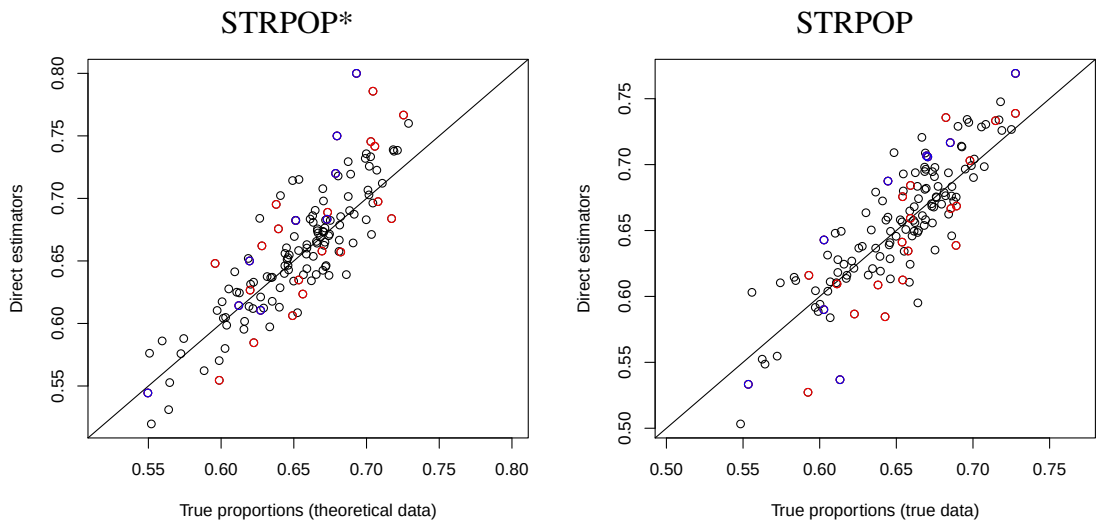

Figure 4: Direct estimators and true values of district proportions for theoretical and true data. Districts with sample sizes smaller than 200 appear in red and those for districts smaller than 100 in blue.

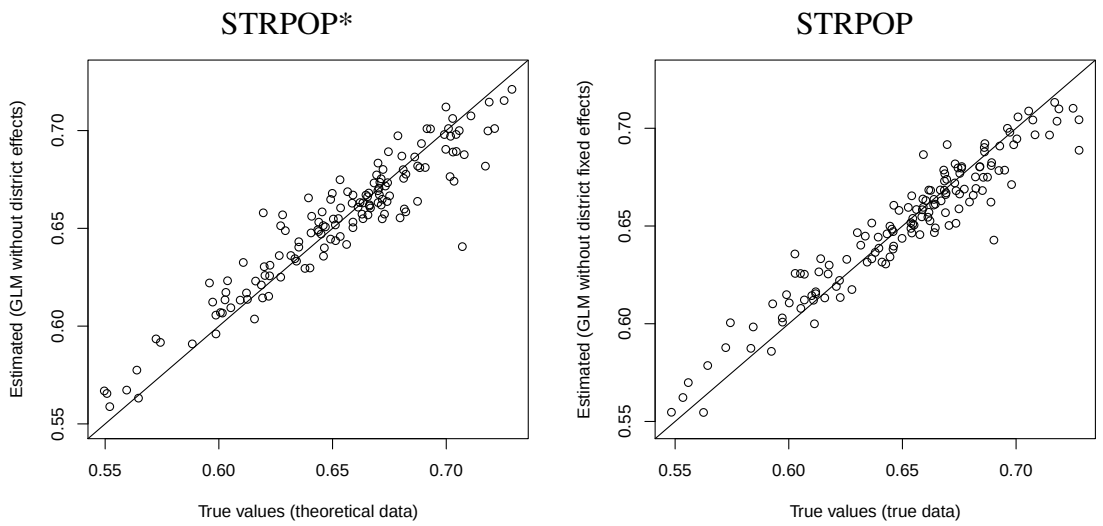

Figure 5: Plug-in estimators based on GLM without district effects against true values for theoretical data STRPOP* and for true data STRPOP.

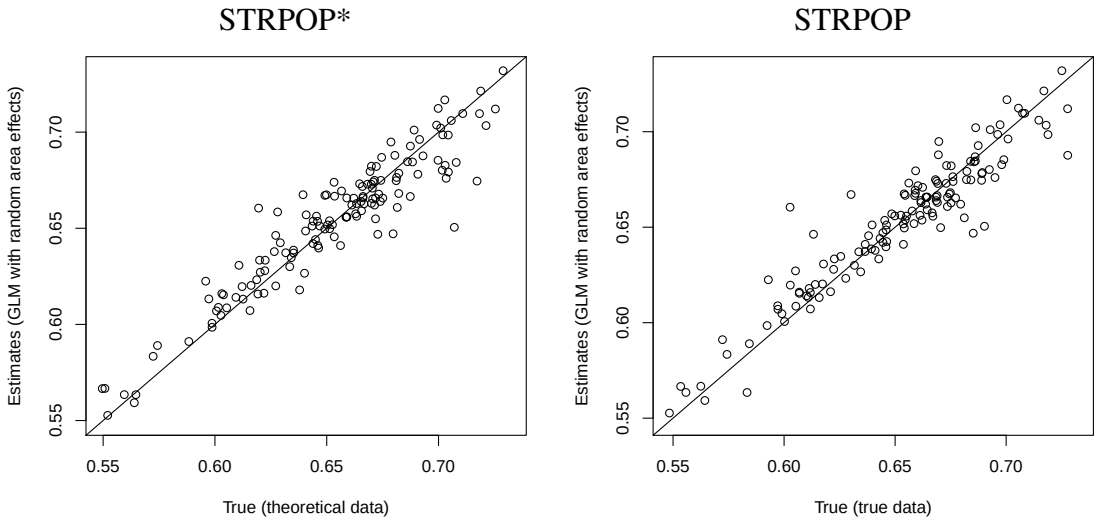

Figure 6: Plug-in estimators based on GLMM that includes random district effects for theoretical and true data.

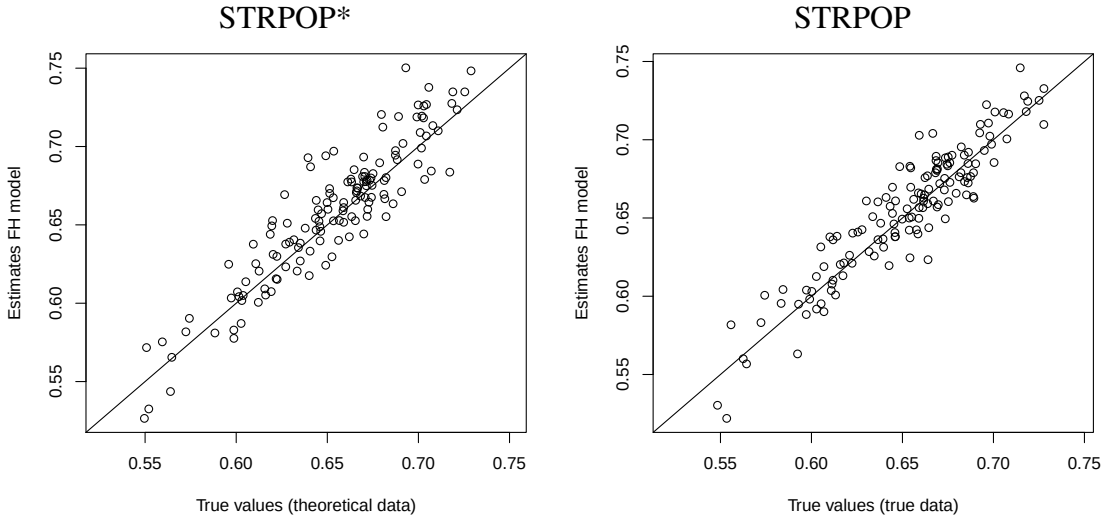

Figure 7: FH estimators for theoretical and true data.

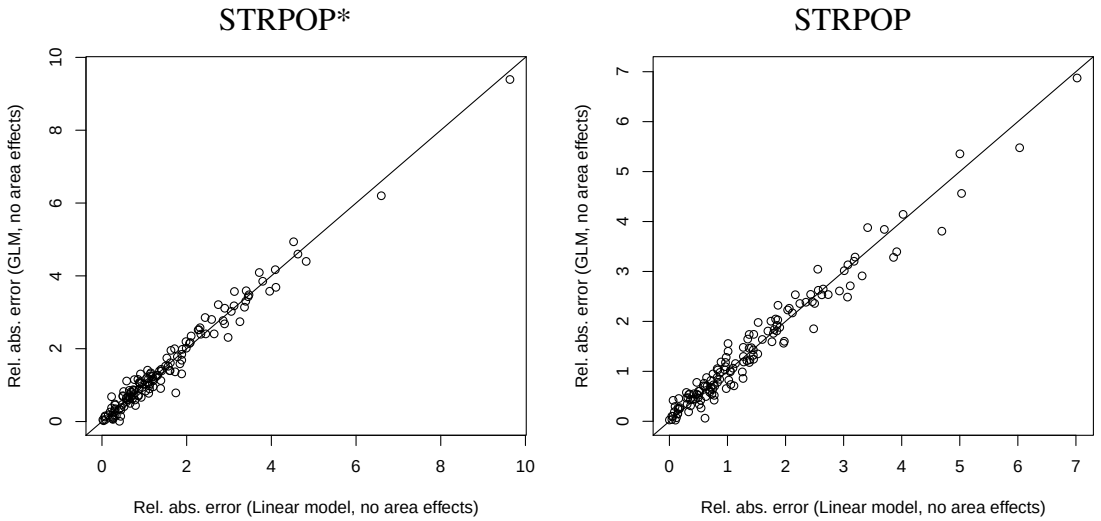

Figure 8: Relative error: LM v.s. GLM without area effects for theoretical and true data.

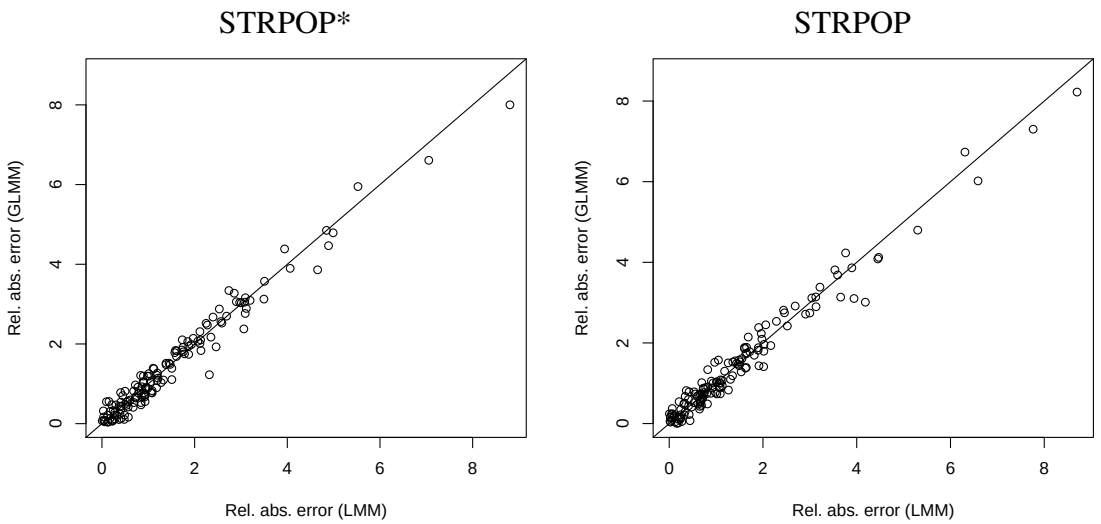

Figure 9: Relative error: LMM v.s. GLMM for theoretical and true data.

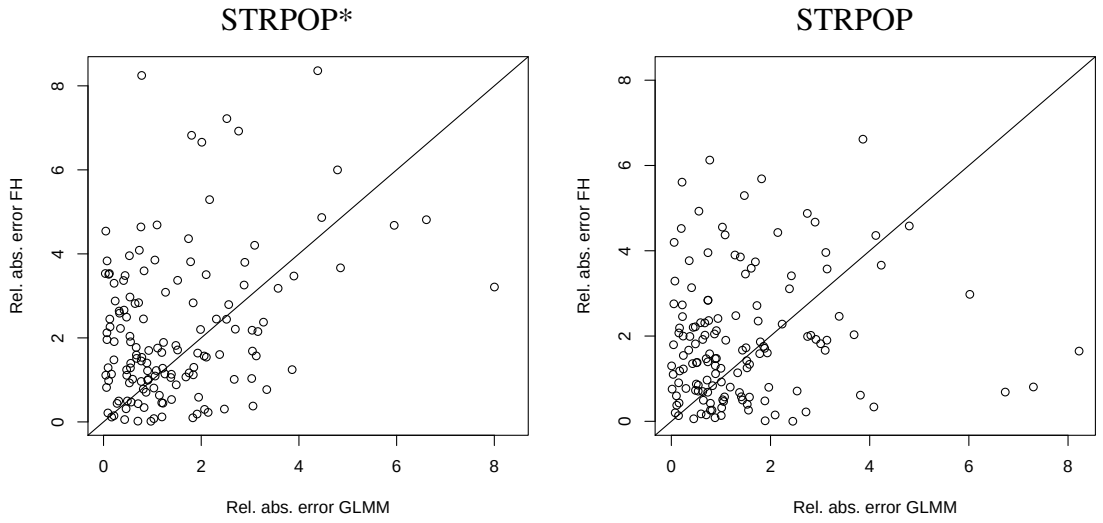

Figure 10: Relative error: GLM with random area effects vs. FH model for theoretical and true data.

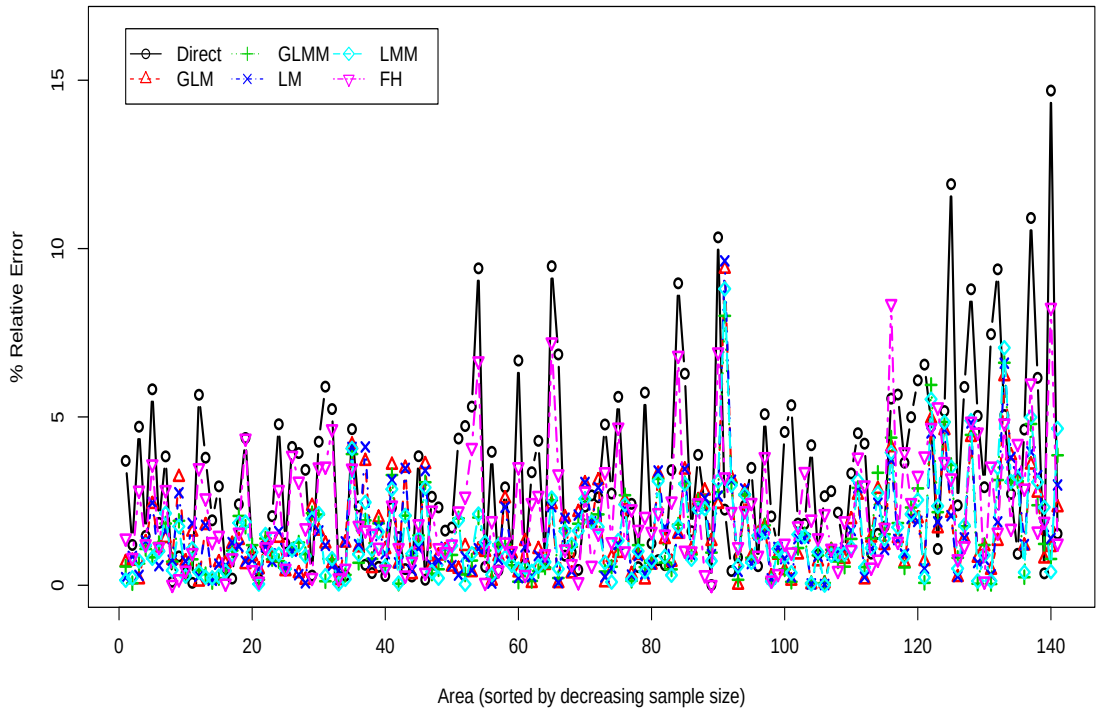

Figure 11: Percent relative errors of estimators based on the direct,GLM, GLMM, LM, LMM and FH for each area, with districts sorted by decreasing sample sizes.

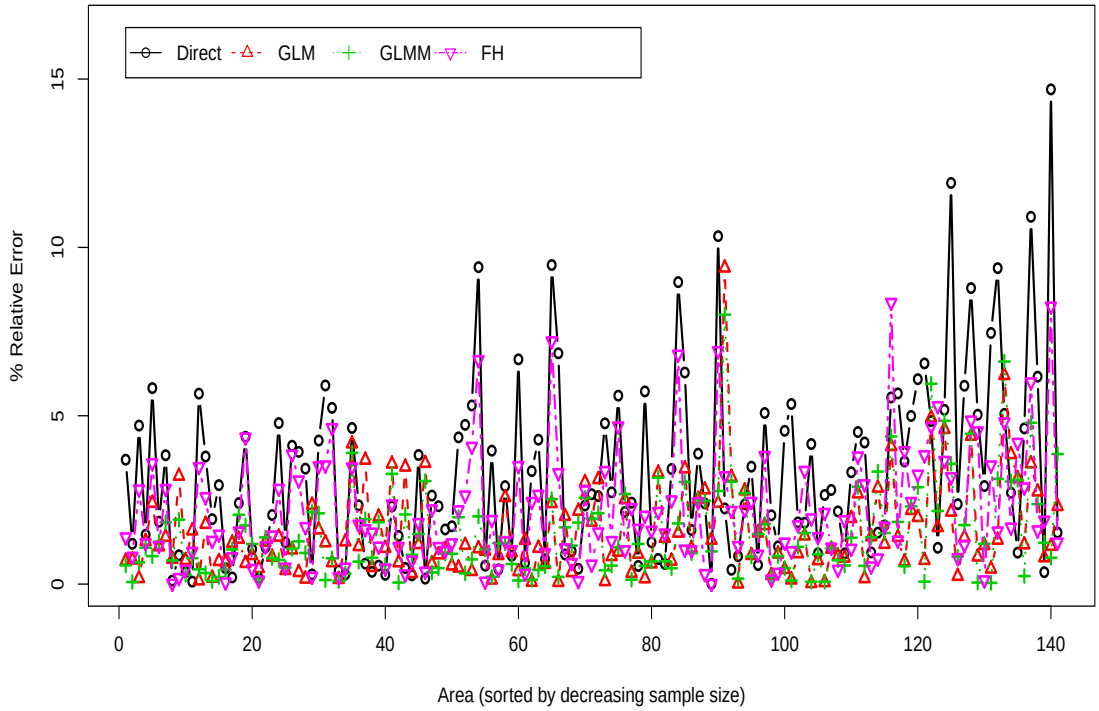

Figure 12: Percent relative errors of estimators based on the Direct, GLM, GLMM and FH for each area, with districts sorted by decreasing sample sizes.

6 Model-based simulation experiment

In the following simulation experiment, we consider a fixed sample but we analyze the properties of estimators with respect to the model distribution, that is, for all possible vectors of values of the population units (model-based simulation). This allows us to study the properties of the considered estimators under a controlled situation in which we know the distribution of the population vector (the model). For each Monte Carlo simulation, the populations of Y_{di} values are generated the same as described in previous section. The sample (with the same indices over all Monte Carlo replicates) is generated also as described before.

Results are depicted in Figures 13-18. Figure 13 plots the means of true values and estimates across Monte Carlo simulations for all districts. The FH estimate is plotted in Figure 14 together with estimates based on generalized linear models. For a given district, the difference between the mean of the Monte Carlo values of an estimator and the mean of the true proportions is exactly the bias of this estimator. This figure shows that the means of model-based estimates are tracking the means of true values better than direct estimates under this setup.

For a given district and estimator, the absolute relative bias (ARB) has been calculated

as the absolute bias divided by the mean of the true values. Percent ARBs for all districts are plotted in Figures 15 and 16. Figure 15 shows that direct estimates have large ARBs for many districts. The two synthetic models, LM and GLM, are giving very similar ARBs for all districts and have larger ARBs than direct estimates for several districts. From these two models, the GLM gives just slightly better results in the worst cases. Mixed models accounting for random area effects, LMM and GLMM also provide similar estimators and both have smaller ARBs than direct estimators for most districts, with the GLMM giving slightly better results than the LMM in the worst cases. Figure 16 shows only direct, FH and estimates based on the GLMM. Aver all the districts, we can see that FH estimates are doing better than direct estimates but not as good as GLMM ones.

The relative root mean squared error (RRMSE) of an estimator is calculated by averaging across Monte Carlo replicates the squared differences between the values of the estimator and the true values, then taking squared root and finally dividing by the mean of the true values. Percent RRMSEs are plotted in Figures 17 and 18. Again, we can see that mixed models, LMM and GLMM, provide estimators with much smaller RRMSEs than direct estimators for practically all districts. There is only one district in which these models provide an estimator with significantly larger RRMSE than that of the direct estimator. Since results are averaged over populations and the sample here is kept fixed, this particular subsample is giving a better direct estimator than the ones delivered by the models, but this does not happen when averaging over all samples, see Figure 23. There is another district where the mixed models are giving an estimate with a larger RRMSE of about 4%, but the direct estimator does a little worse with a RRMSE of about 5%. Figure 24 shows again that FH estimates are not borrowing so much strength than estimates based on the GLMM, achieving larger RRMSEs for most districts and even larger than those of direct estimators for several districts.

7 Design-based simulation experiment

A similar experiment is carried out, where the population is generated in the same way as before, but in this case it is kept fixed and $L = 250$ different samples are drawn from it. Thus, in this case the properties of the estimators are analyzed under the sampling design (design-based simulations). Still, the fixed population is generated from the GLM and therefore we keep the control on the true distribution that generates the population data.

Results are depicted in Figures 19-24. Figure 19 shows the true proportions together with the means over the L samples of direct estimates and estimates based on GLM, GLMM, LM and LMM for each district in the x axis. All estimates seem to track acceptably well the true values for most districts, with only small deviations shown by model-based estimates for few districts. Figure 20 plots the FH estimates together with the estimates based on the GLM and GLMM. Again, FH estimates seem to behave satisfactorily in terms of bias, showing also slight biases for few districts. The absolute relative design-bias is depicted in Figures 21 and 22. Direct estimates are design-unbiased, and this is clearly seen in these two plots. Synthetic models, LM and GLM, appear to have larger design-biases than the models with random district effects, LMM and GLMM. The latter models get very similar relative biases, but the GLMM gets a slightly smaller ARE

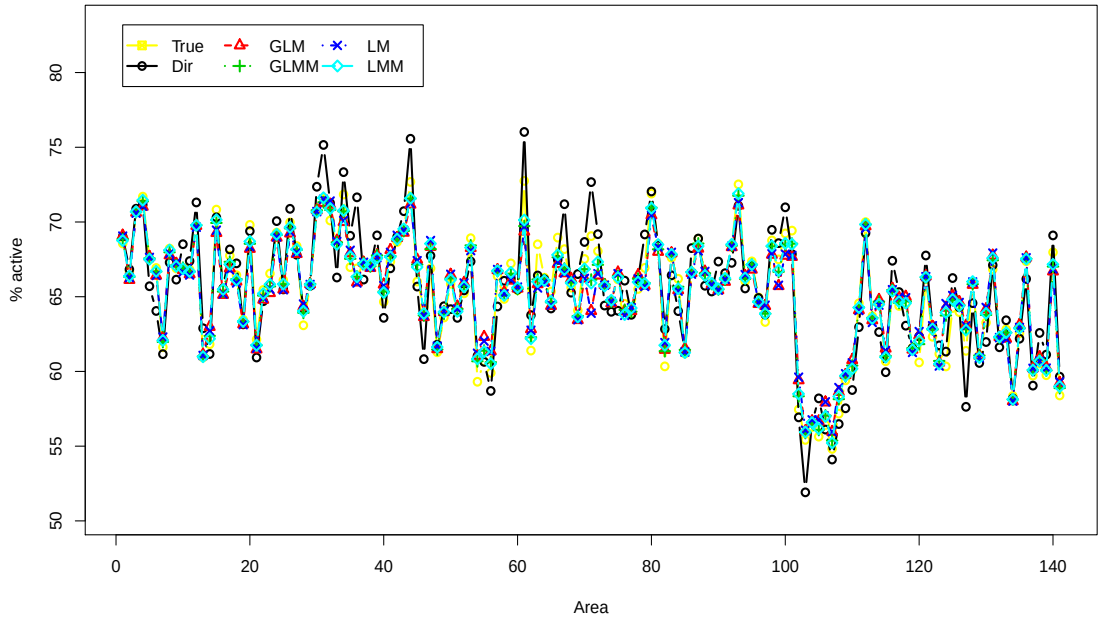

Figure 13: (Model-based simulation) Mean of true values, direct estimates and estimates based on the GLM, GLMM, LM and LMM for each district.

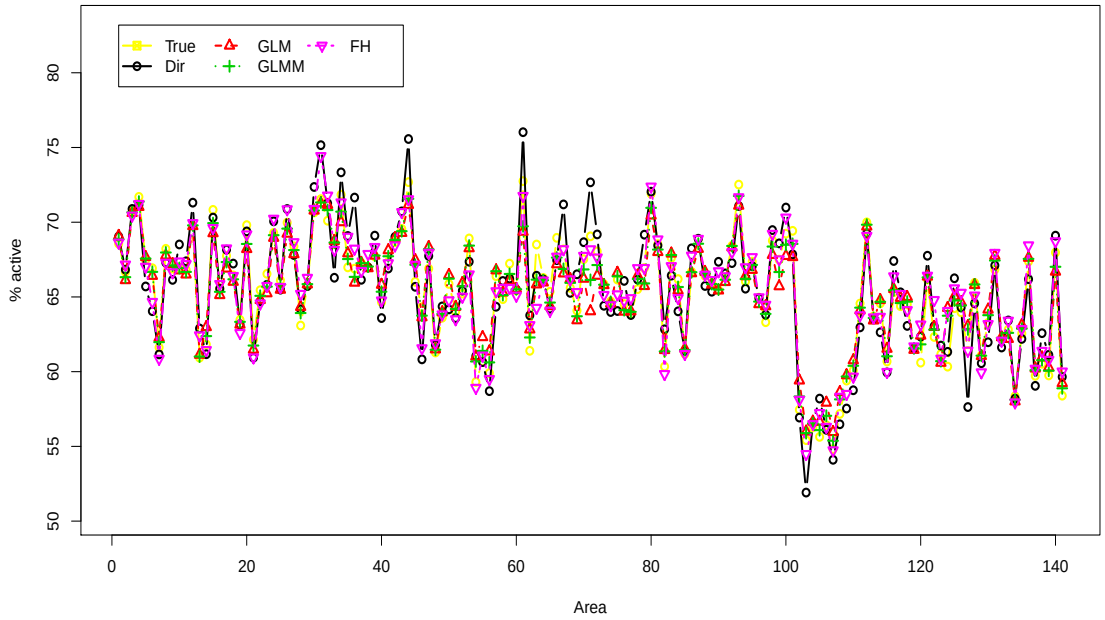

Figure 14: (Model-based simulation) Mean of true values, direct and FH estimates, and estimates based on the GLM and GLMM for each district.

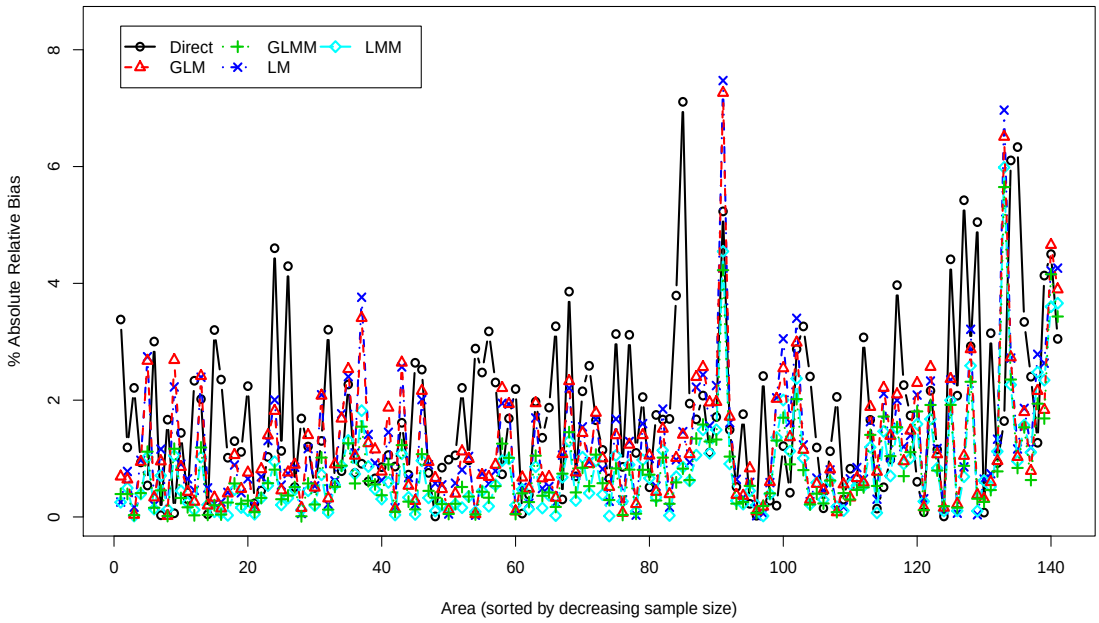

Figure 15: (Model-based simulation) Percent Absolute Relative Bias of direct estimators and estimators based on the GLM, GLMM, LM and LMM for each district, with districts sorted by decreasing sample sizes.

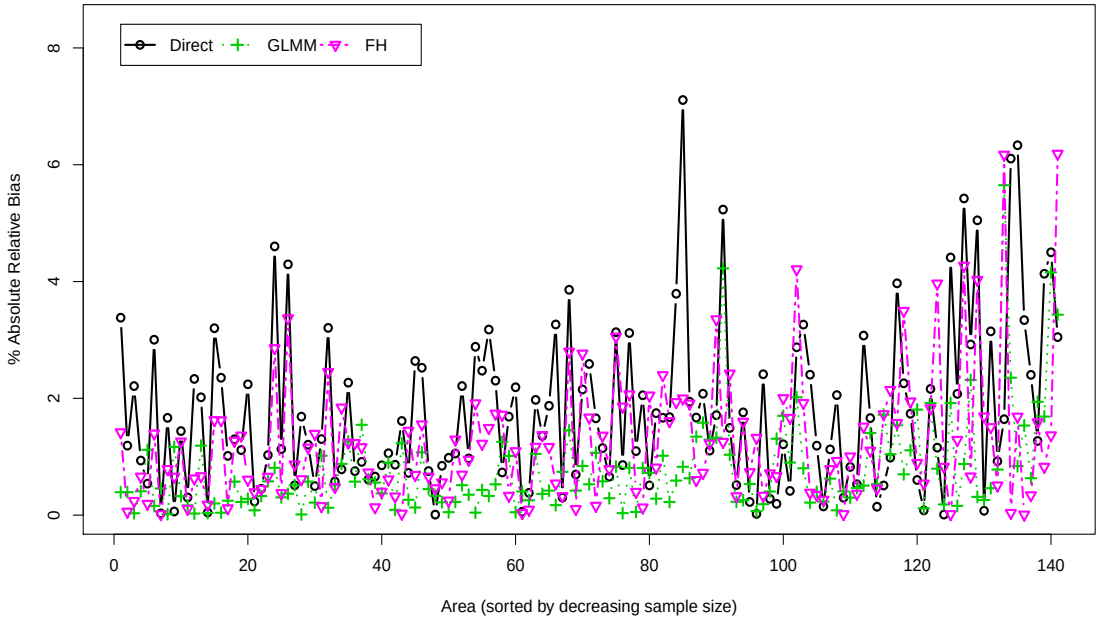

Figure 16: (Model-based simulation) Percent Absolute Relative Bias of direct and FH estimates, and estimates based on the GLM and GLMM for each district, with districts sorted by decreasing sample sizes.

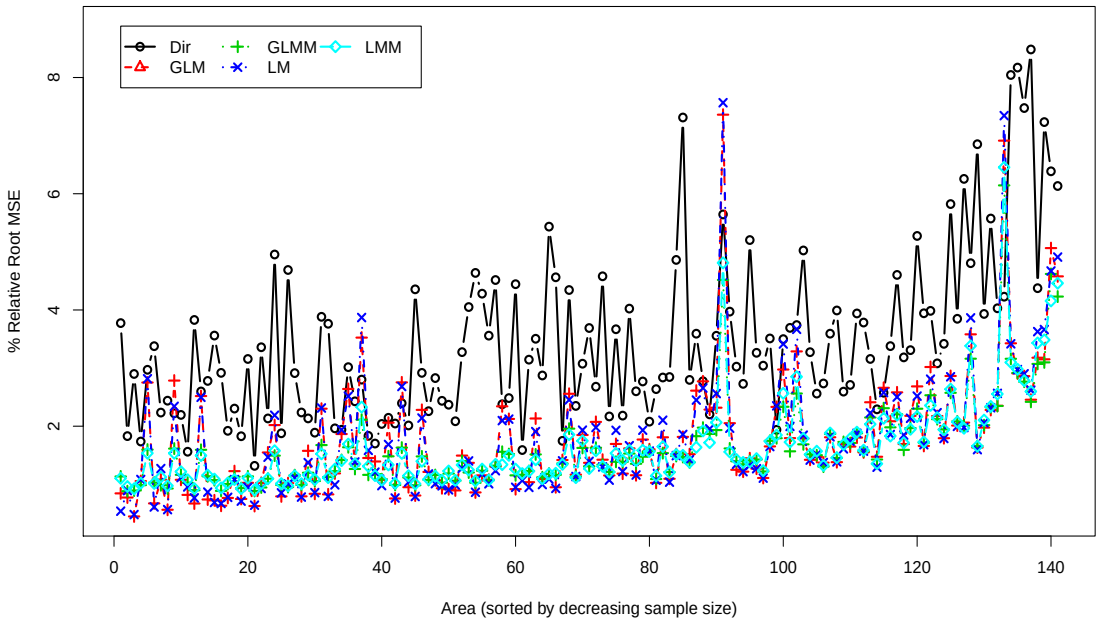

Figure 17: (Model-based simulation) Percent relative root mean squared error of direct estimators and estimators based on the GLM, GLMM, LM and LMM for each district, with districts sorted by decreasing sample sizes.

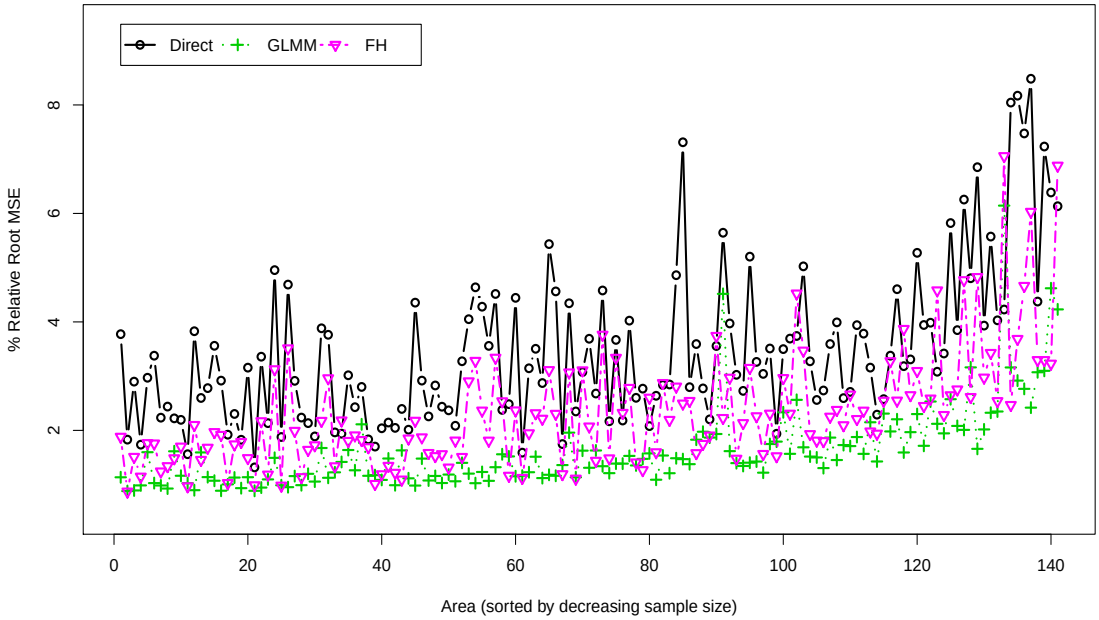

Figure 18: (Model-based simulation) Percent relative root mean squared error of direct and FH estimators, and estimators based on the GLMM for each district, with districts sorted by decreasing sample sizes.

in the worst case, and keeps all absolute relative biases smaller than 5% for all districts.

When it comes to compare the true mean squared errors (MSEs) under the design-based setup, in the case of direct estimators, the variance is the main term contributing to the mean squared error. In the case of model-based estimators, it is the design-bias already observed above the term that mostly contributes to the MSEs. In fact, as we can see in Figures 23 and 24 showing the percent relative root mean squared errors (RRMSEs), direct estimators are considerably more inefficient than estimators based on LMM and GLMM for all districts except one. In that district, models give worse RRMSE but still for the GLMM this error does not exceed 5%. For practically all districts, mixed models accounting for random area effects are reducing the error to a great extent. Synthetic models, LM and GLM, provide more accurate estimators than the direct ones for most districts, but in some of them are doing worse, and especially worse for the mentioned district. The activity in this district seems not to be fully explained by the available auxiliary variables and it seems to be an outlier as compared with the other districts, in which models explain adequately the activity. A possible solution for this district is to include a single fixed effect in the model only for it, in which case models will provide estimators for that district that will be close to its direct estimator. Figure 24 shows the RRMSEs for direct estimators, FH estimators and the estimators provided by the best model, GLMM. FH estimates also improve the error of direct estimators for all districts except few of them, but the reductions in error achieved by the estimators obtained from the GLMM are greater. These conclusions confirm the answers to the questions given when analysing only one sample in Section 5.

8 Design-based simulation based on true population

Section 7 studies the behavior of the estimators under optimal conditions in which the data generating process is known, because the theoretical population STRPOP* was generated from the fit of the GLM with fixed effects. Now we analyse when the true data STRPOP is taken as the population. Thus, similarly as in Section 7 but taking the original data set STRPOP as the fixed population, a design-based simulation study is performed by drawing $L = 500$ different samples from it. Results are plotted in Figures 25-30. Note that plots are similar to those based on the theoretical population STRPOP* shown in Section 7. This is not surprising since we already saw that the model used to generate STRPOP* was fitting adequately the data. The plots confirm all our previous comments and show that estimators based on the models with random district effects are performing considerably better than the other estimators overall districts. The GLMM gives just slightly better estimates than the LMM. Finally, FH estimators are not as efficient as the other model-based estimators. There is only one district in which the estimator based on the GLMM has a larger RRMSE than the direct estimator. The activity in this district seems not to be well explained by the considered models with the given auxiliary variables and a deeper analysis is needed for this district. However, note that for this district, the RRMSE of the estimator based on the GLMM is below 5%.

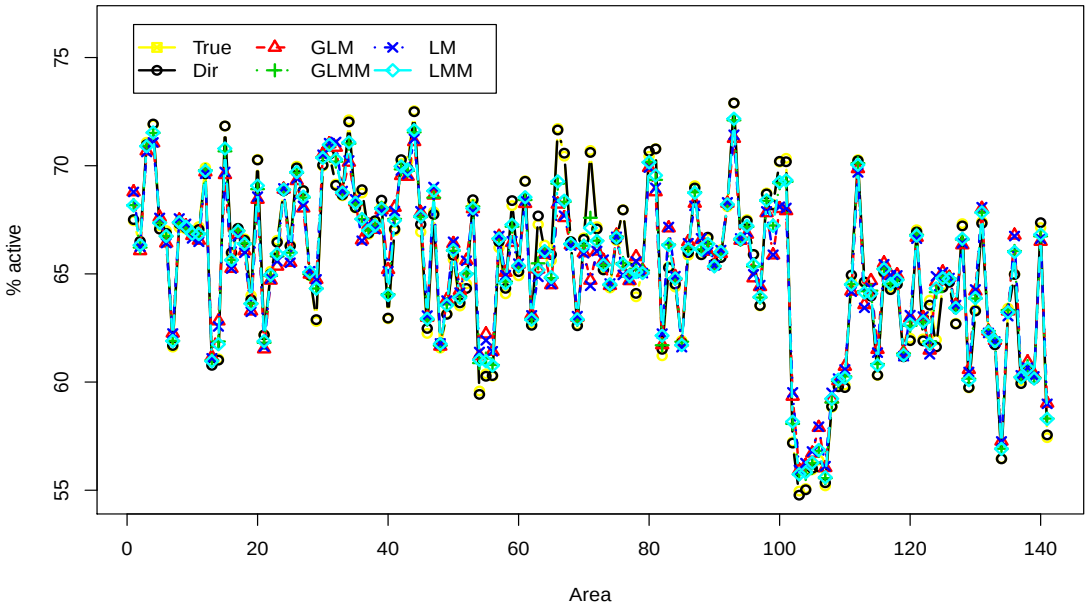

Figure 19: (Design-based simulation) True values and means of direct estimates and estimates based on the GLM, GLMM, LM and LMM for each district.

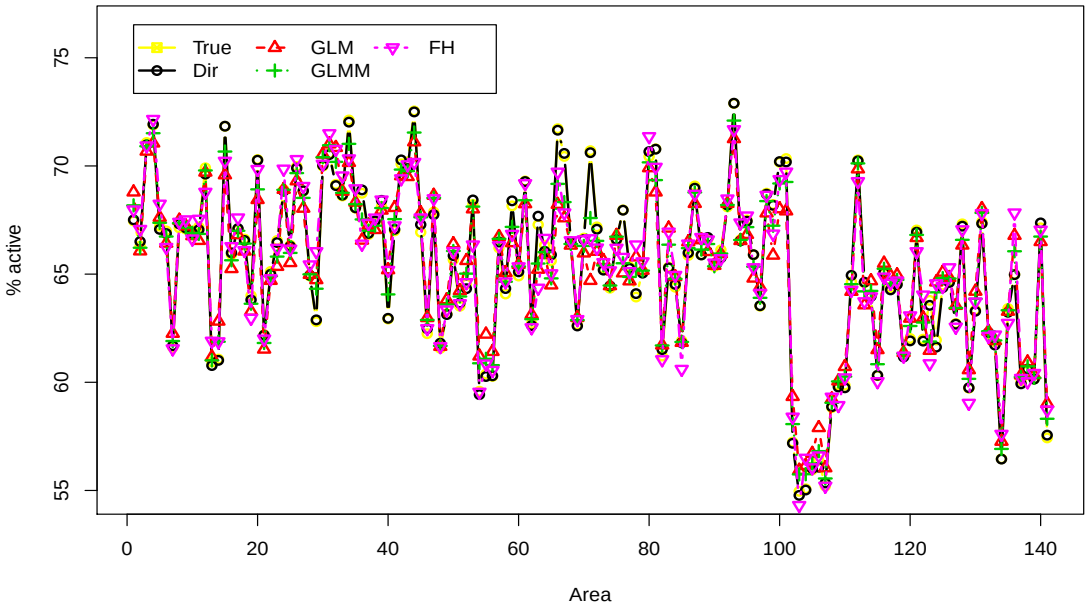

Figure 20: (Design-based simulation) True values and means of direct and FH estimates, and estimates based on the GLM and GLMM for each district.

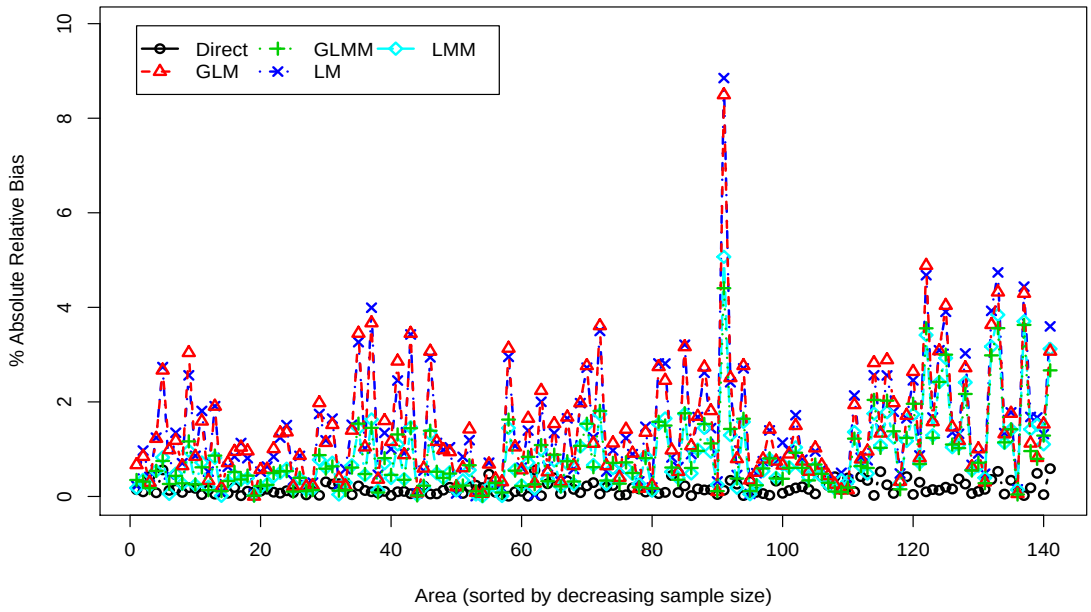

Figure 21: (Design-based simulation) Percent Absolute Relative Bias of direct estimators and estimators based on the GLM, GLMM, LM and LMM for each district, with districts sorted by decreasing sample sizes.

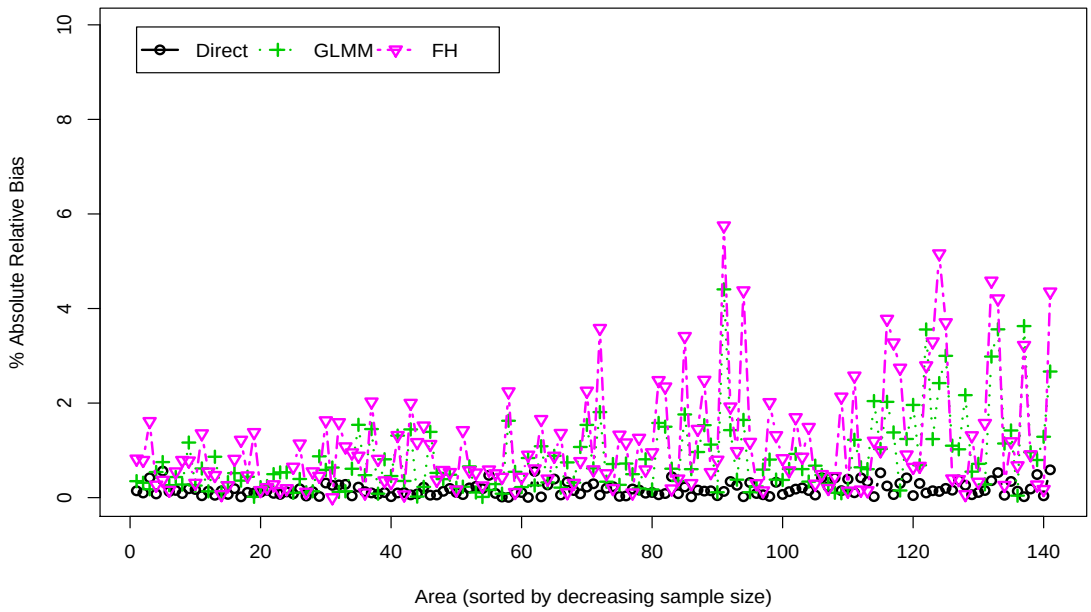

Figure 22: (Design-based simulation) Percent Absolute Relative Bias of direct and FH estimates, and estimates based on the GLM and GLMM for each district, with districts sorted by decreasing sample sizes.

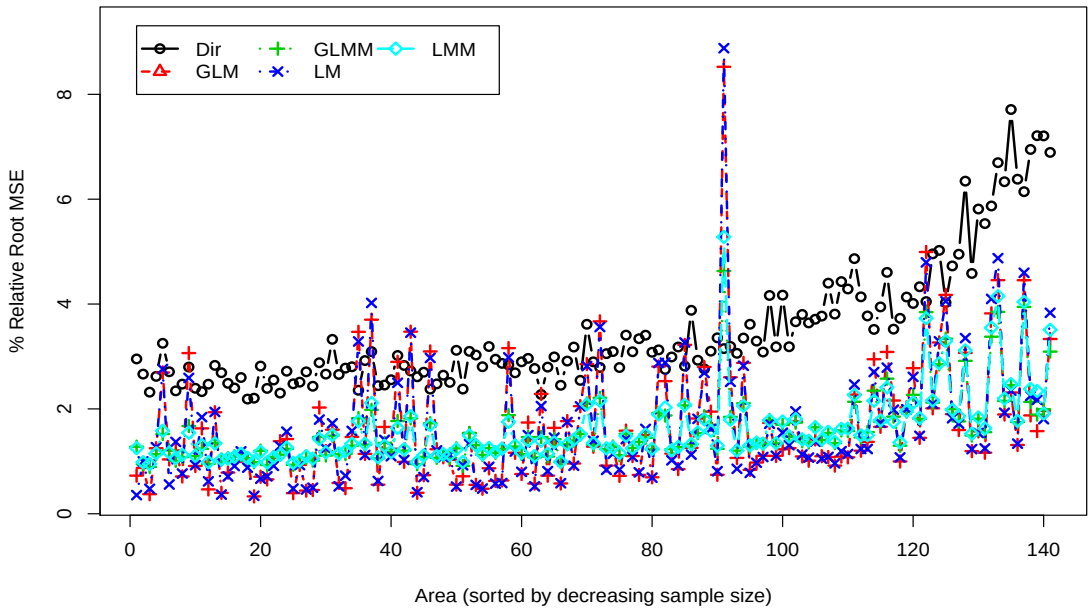

Figure 23: (Design-based simulation) Percent relative root mean squared error of direct estimators and estimators based on the GLM, GLMM, LM and LMM for each district, with districts sorted by decreasing sample sizes.

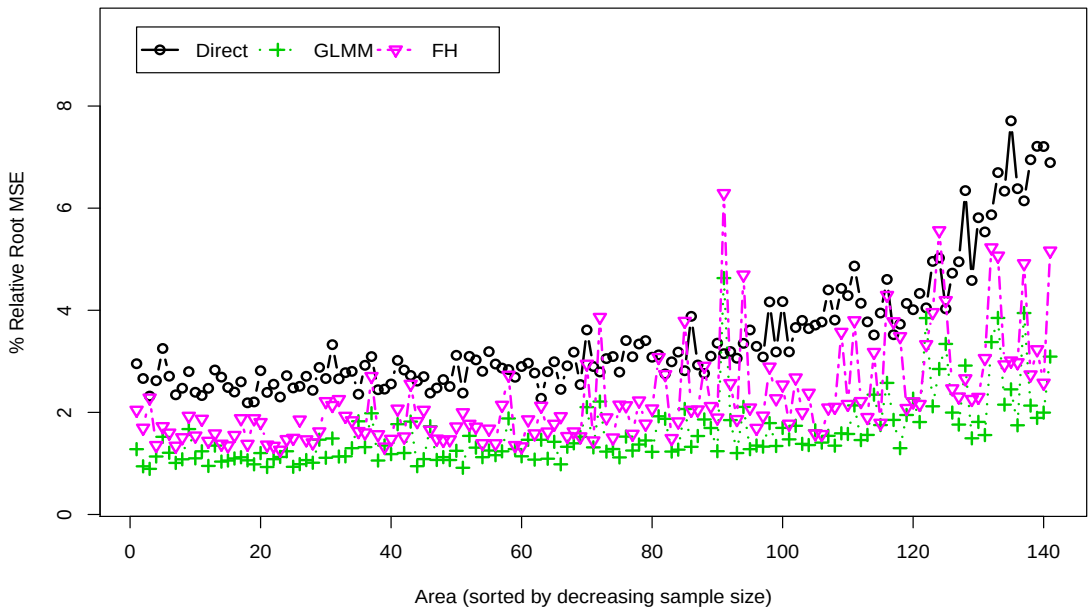

Figure 24: (Design-based simulation) Percent relative root mean squared error of direct and FH estimators, and estimators based on the GLMM for each district, with districts sorted by decreasing sample sizes.

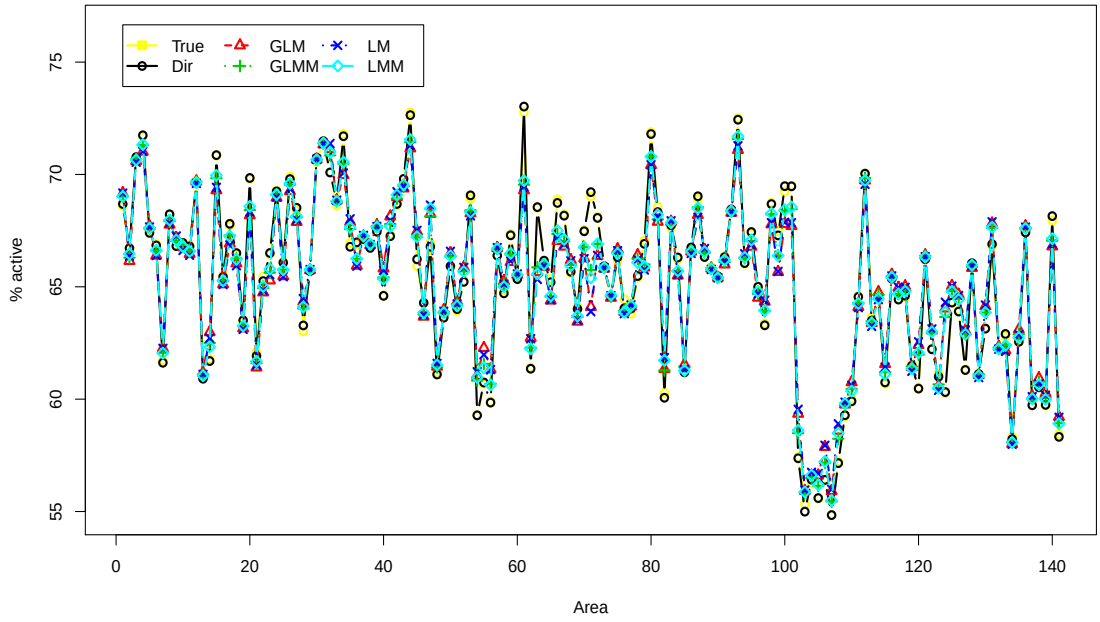

Figure 25: (Design-based simulation with true data) True values and means of direct estimates and estimates based on the GLM, GLMM, LM and LMM for each district.

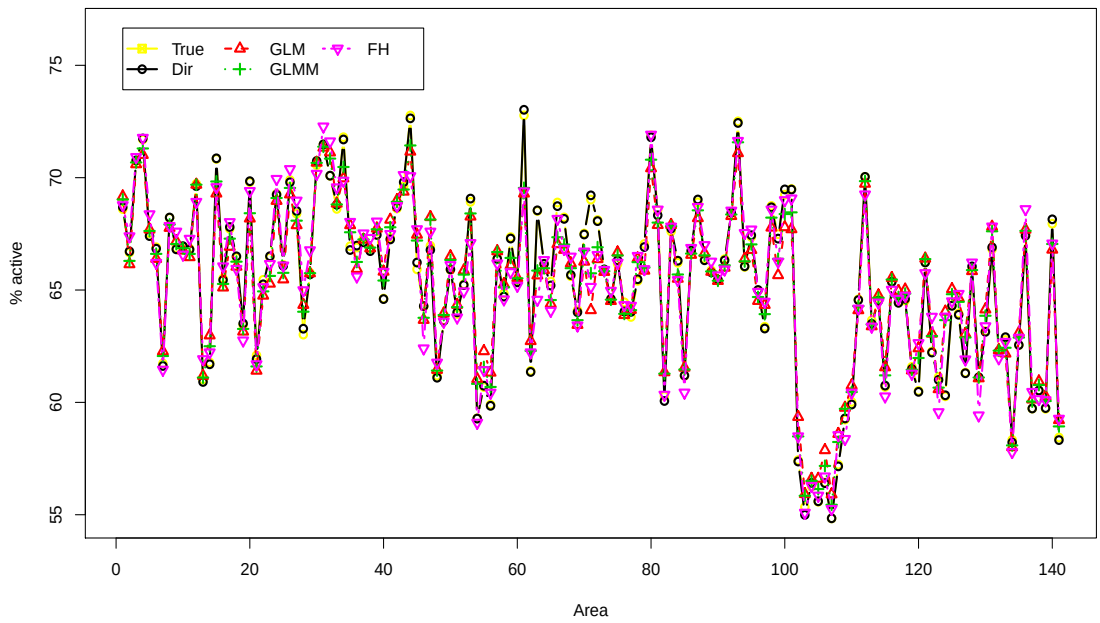

Figure 26: (Design-based simulation with true data) True values and means of direct and FH estimates, and estimates based on the GLM and GLMM for each district.

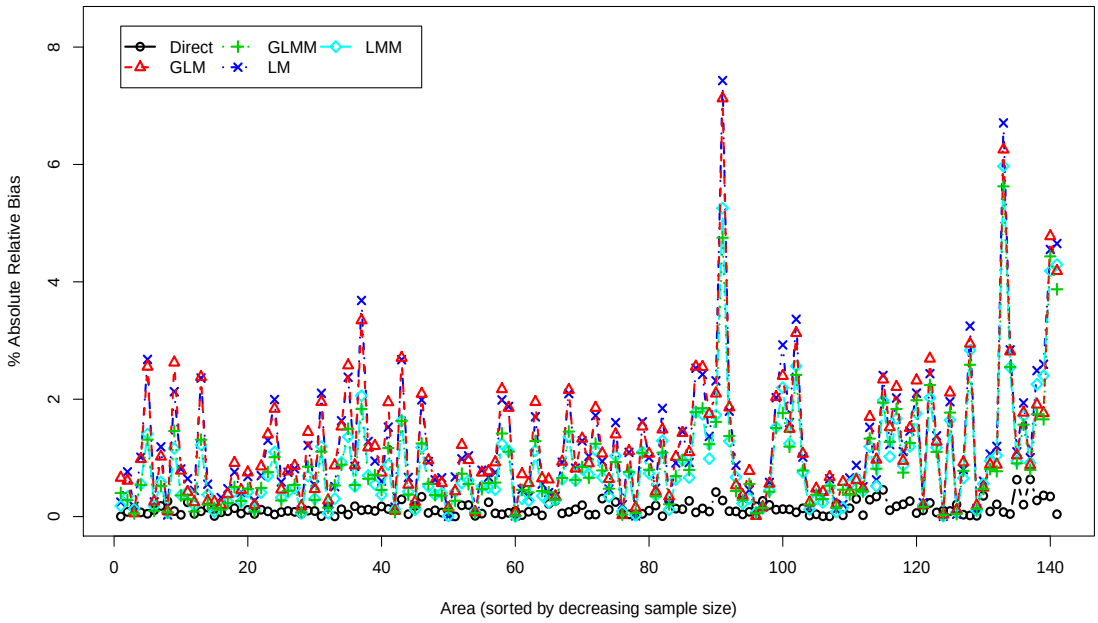

Figure 27: (Design-based simulation with true data) Percent Absolute Relative Bias of direct estimators and estimators based on the GLM, GLMM, LM and LMM for each district, with districts sorted by decreasing sample sizes.

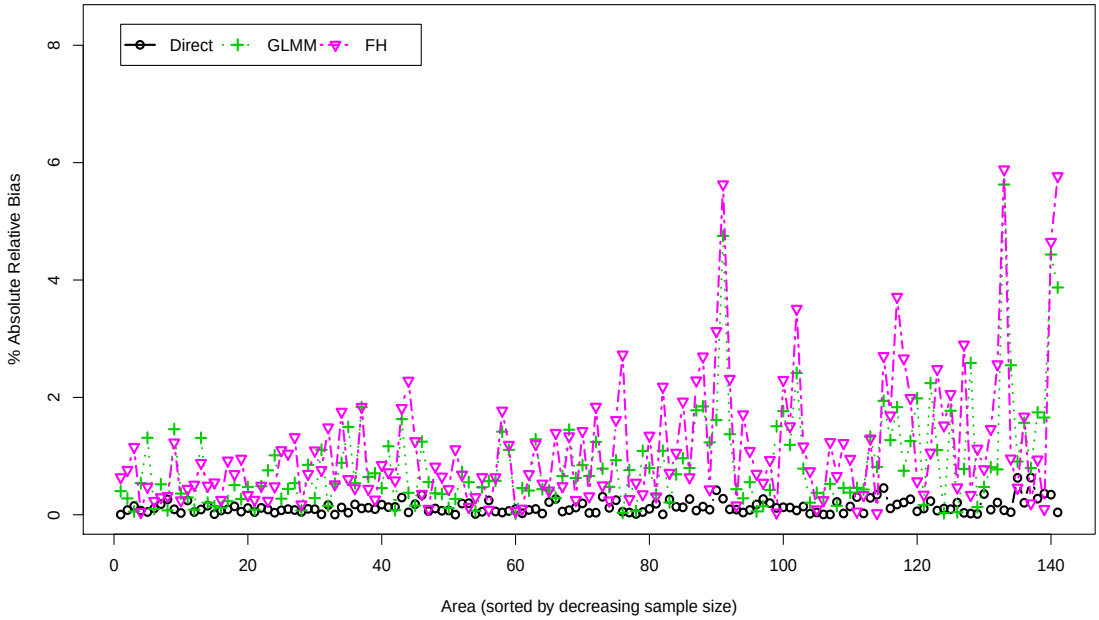

Figure 28: (Design-based simulation with true data) Percent Absolute Relative Bias of direct and FH estimates, and estimates based on the GLM and GLMM for each district, with districts sorted by decreasing sample sizes.

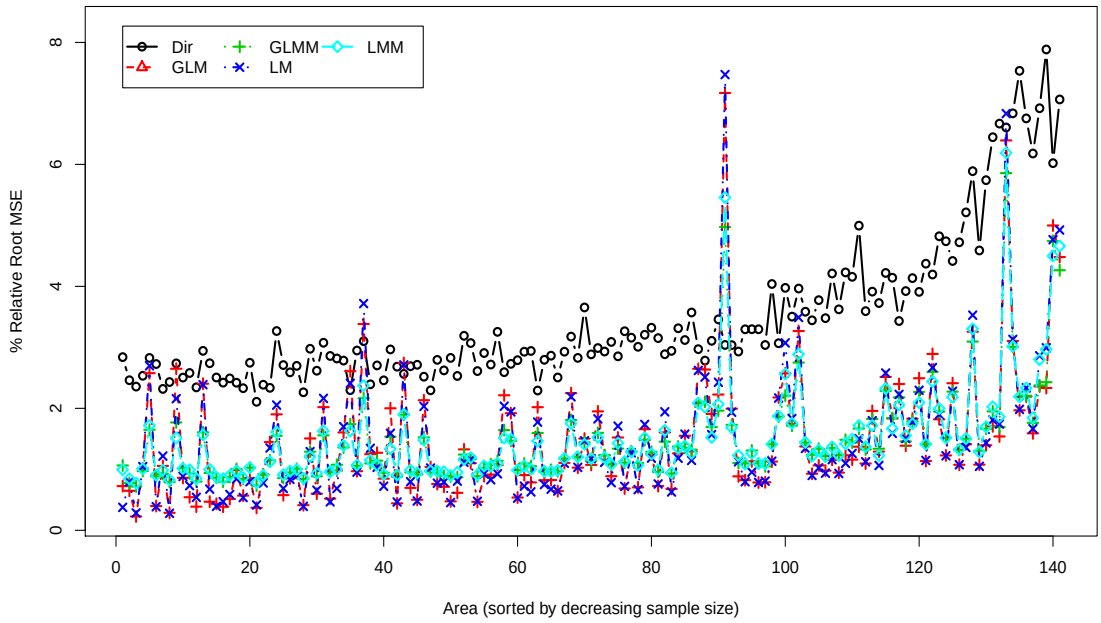

Figure 29: (Design-based simulation with true data) Percent relative root mean squared error of direct estimators and estimators based on the GLM, GLMM, LM and LMM for each district, with districts sorted by decreasing sample sizes.

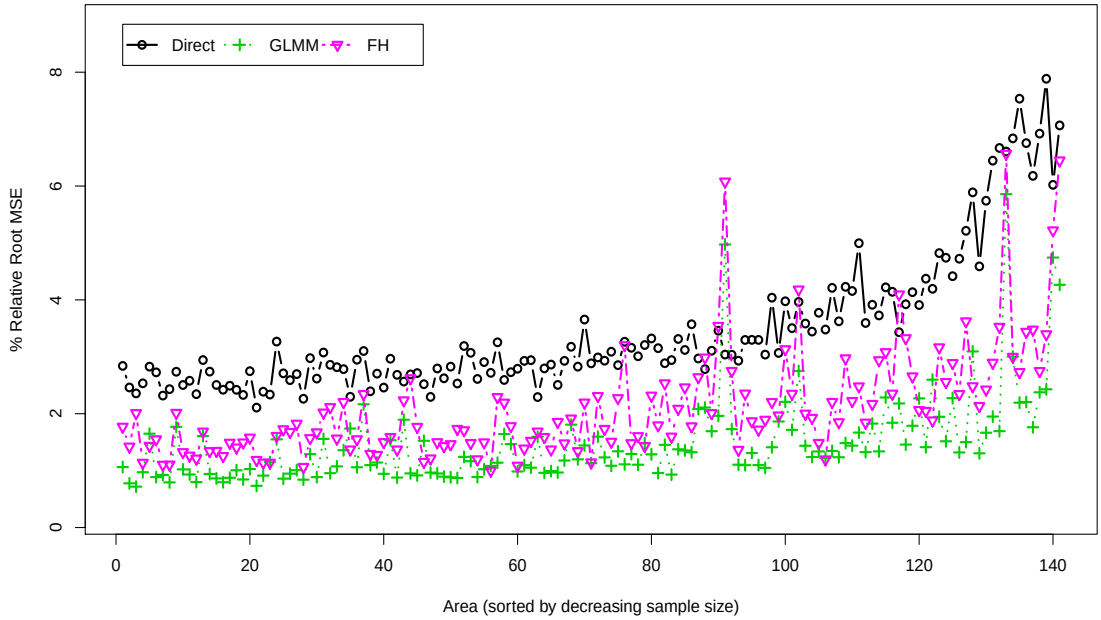

Figure 30: (Design-based simulation with true data) Percent relative root mean squared error of direct and FH estimators, and estimators based on the GLMM for each district, with districts sorted by decreasing sample sizes.

9 Final remarks

In this section we discuss all the results obtained above. We have seen that direct estimators are considerably less efficient than estimators based on models with random district effects. Concretely, those based on the GLMM are slightly better than the ones based on the LMM. This comment is true except for one district, which is the one identified as "1724" and which belongs to Strata coded "SG00". In order to understand the problem in that particular district, we take a look at the estimated random effects of the GLMM. Figures 31 and 32 show respectively the histogram and the Q-Q plot of estimated random effects of the districts with and without district "1724". The asymmetry of the histogram and the Q-Q plot is caused by this sole district, because when excluding it from the plots, the histogram becomes symmetric and fits well a normal distribution, and the quantiles in the Q-Q plot without district "1724" fit a straight line. This confirms that, removing this problematic district, the assumption of normality for the random effects is approximately correct. Thus, district "1724" is certainly an outlying district in which the models with the considered variables are not able to explain the activity.

In the previous simulation studies, we had selected a subset of the original data from the Structural Survey, in which the districts with smallest sample sizes were removed in order to have reliable "true values". Then smaller samples were drawn from the remaining districts, taking sample sizes such that the smallest ones were similar to those of the smallest districts that were removed. Under this situation, we have found that the best model for estimation of the proportions of active people is the GLMM that includes random area effects. Once we have selected the best model, we fitted it to the whole data set from the Structural Survey. Figure 33 shows the final estimators obtained from the selected GLMM together with direct estimators for all districts. Figure 34 shows the same results with districts sorted by decreasing sample sizes. See the great similarities between the two sets of estimates for the districts with larger sample sizes (those on the left-hand side) and the small discrepancies for the districts with smaller sample sizes (on the right-hand side). Moreover, Figure 35 shows a line plot of the the two sets of estimates. The fact that points lie around the line and are scattered at both sides of the line indicates no significant systematic bias of estimates based on the GLMM under the design-based setup. Finally, Table 8 shows the estimated regression coefficients in the GLMM model (fixed effects only).

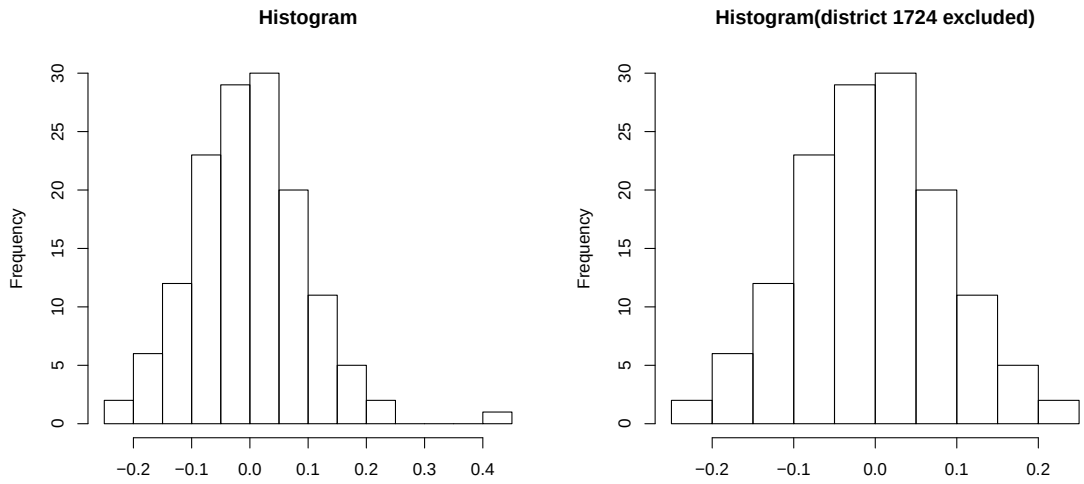

Figure 31: Histogram of estimated random effects using the GLMM model for all districts (left) and without district “1724” (right).

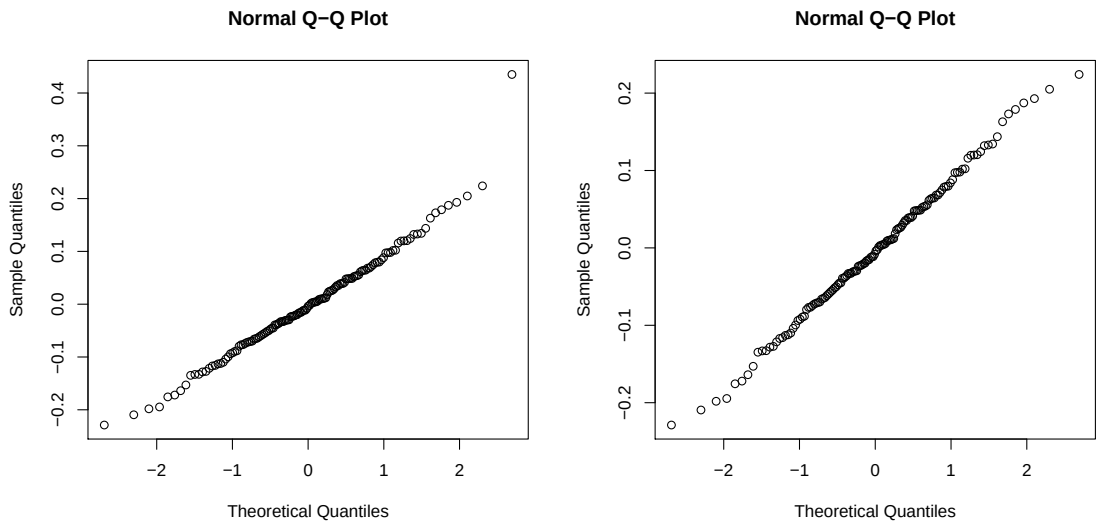

Figure 32: Q-Q plot of estimated random effects using the GLMM for all districts (left) and without district “1724” (right).

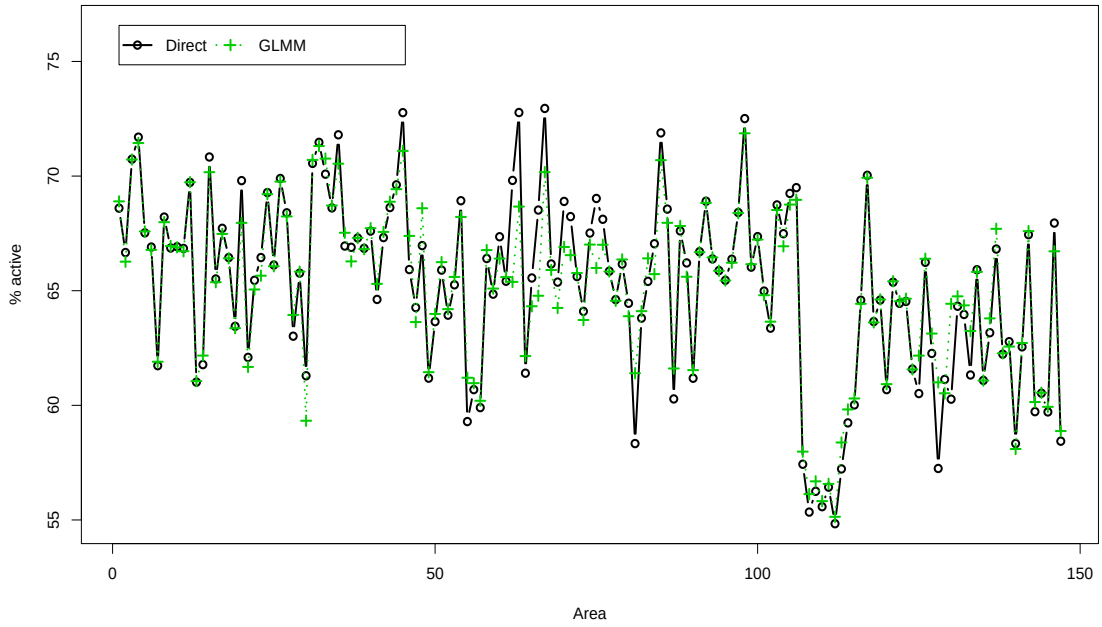

Figure 33: Estimated percentage of actives using the direct estimators and the ones based on the GLMM for the Structural Survey data with all districts.

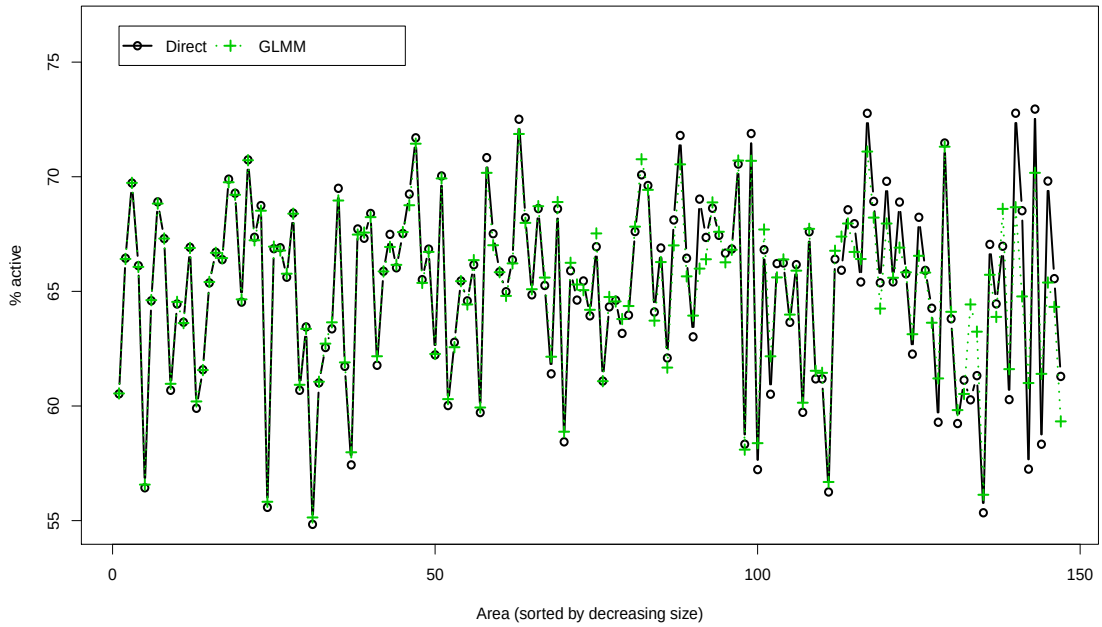

Figure 34: Estimated percentage of actives using the direct estimators and the ones based on the GLMM for the Structural Survey data with all districts, with districts sorted by decreasing sample sizes.

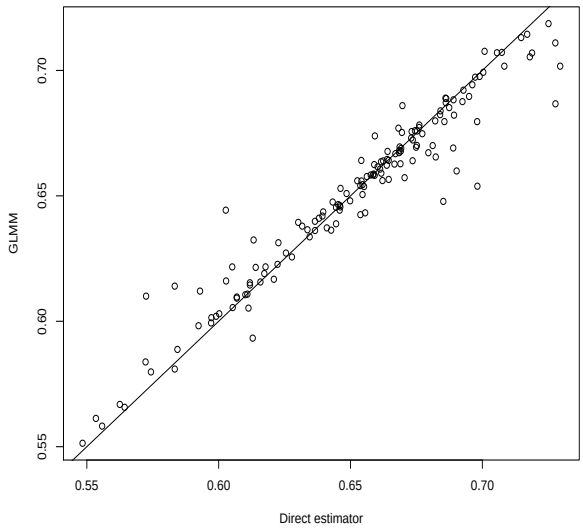

Figure 35: Estimated proportions of actives using the direct estimators and the GLMM model for the Structural Survey data with all districts.

Table 8: Regression coefficients ($\hat{\beta}$)

Name	Intercept	Strata1	age group			gender
Category		1	2	3	4	2
$\hat{\beta}$	0.0198	-0.0989	0.0717	-2.2492	-3.6653	-0.1049
Name	civil status			nationality	secondary residence	Household Size
Category	2	3	4	2	2	2
$\hat{\beta}$	0.4859	0.0003	0.2961	-0.1927	-0.6367	-0.0951
Name	Household Size			Income		
Category	3	4	5	2	3	4
$\hat{\beta}$	-0.2729	-0.4103	0.0027	2.1298	3.5943	4.3477
Name	age group*gender			gender*civil status		
Category	2(2)	3(2)	4(2)	2(2)	2(3)	2(4)
$\hat{\beta}$	0.0593	0.6551	0.7941	-0.8113	-0.6487	-0.3753

Appendix 1: Considered estimators

Let U be the population of size N, composed of D non-overlapping areas $U_1, \ldots, U_D$, of sizes $N_1, \ldots, N_D$ with $N = \sum_{d=1}^{D} N_d$. Let s be a sample of size n drawn from U and s_d

the subsample from area d of size n_d, $d = 1,\ldots,D$, where $n = \sum_{d=1}^{D} n_d$. Let $\bar{s}_d = U_d - s_d$ denote the complement of the sample from area d. Let Y_{di} be the target variable for unit i in area d. The target parameters are the true proportions

$$P_d = N_d^{-1} \sum_{i=1}^{N_d} Y_{di}, \quad d = 1,\ldots,D.$$

We consider estimators based on assuming different models. The following subsections list the assumed models in each case and the estimators derived from those models.

Generalized Linear Model

Consider the following generalized linear model (GLM) for the target variables:

$$Y_{di} \overset{ind.}{\sim} \text{Bern}(p_{di}), \quad p_{di} = \frac{\exp(\mathbf{x}'_{di}\boldsymbol{\beta})}{1+\exp(\mathbf{x}'_{di}\boldsymbol{\beta})}, \quad i = 1,\ldots,N_d, \quad d = 1,\ldots,D.$$

The best estimator under this model, which minimizes the mean squared error and is unbiased, is given by

$$\hat{P}_d^{GLM} = \frac{1}{N_d}\left(\sum_{i\in s_d} Y_{di} + \sum_{i\in \bar{s}_d} \hat{p}_{di}^{GLM}\right), \quad d = 1,\ldots,D.$$

where $\hat{p}_{di}^{GLM} = \exp(\mathbf{x}'_{di}\hat{\boldsymbol{\beta}})/\{1+\exp(\mathbf{x}'_{di}\hat{\boldsymbol{\beta}})\}$ are the probabilities predicted through the GLM fit.

Generalized Linear Mixed Model

Consider that the population variables Y_{di} follow the generalized linear mixed model (GLMM) given by

$$\begin{aligned} &Y_{di}|u_d \overset{ind.}{\sim} \text{Bern}(p_{di}), \\ &p_{di} = \frac{\exp(\mathbf{x}'_{di}\boldsymbol{\beta}+u_d)}{1+\exp(\mathbf{x}'_{di}\boldsymbol{\beta}+u_d)}, \quad u_d \overset{iid}{\sim} N(0,\sigma_u^2), \quad i = 1,\ldots,N_d, \quad d = 1,\ldots,D. \end{aligned}$$

The best predictor under this model, which minimizes the mean squared error and is unbiased, is given by

$$\hat{P}_d^{BP} = \frac{1}{N_d}\left\{\sum_{i\in s_d} Y_{di} + \sum_{i\in \bar{s}_d} E(Y_{di}|\mathbf{y}_{ds})\right\}, \quad d = 1,\ldots,D,$$

where $\mathbf{y}_{ds}$ is the sample data from area d. The expected value $E(Y_{di}|\mathbf{y}_{ds})$ cannot be calculated analytically and Monte Carlo simulation methods are required. Instead, we consider the much simpler plug-in estimator

$$\hat{P}_d^{GLMM} = \frac{1}{N_d}\left(\sum_{i\in s_d} Y_{di} + \sum_{i\in \bar{s}_d} \hat{p}_{di}^{GLMM}\right), \quad d = 1,\ldots,D.$$

where $\hat{p}_{di}^{GLMM} = \exp(\mathbf{x}'_{di}\hat{\boldsymbol{\beta}}+\hat{u}_d)/\{1+\exp(\mathbf{x}'_{di}\hat{\boldsymbol{\beta}}+\hat{u}_d)\}$ are the probabilities predicted through the GLMM fit.

Linear Model

The linear model (LM) assumes that the population variables Y_{di} follow the model

$$Y_{di} = \mathbf{x}'_{di}\boldsymbol{\gamma} + e_{di}, \quad e_{di} \overset{iid}{\sim} N(0, \sigma_e^2), \quad i = 1, \ldots, N_d, \quad d = 1, \ldots, D.$$

The EBLUP under this model is given by

$$\hat{P}_d^{LM} = \frac{1}{N_d}\left(\sum_{i \in s_d} Y_{di} + \sum_{i \in \bar{s}_d} \hat{Y}_{di}^{LM}\right), \quad d = 1, \ldots, D,$$

where $\hat{Y}_{di}^{LM} = \mathbf{x}'_{di}\hat{\boldsymbol{\gamma}}$ are the predicted values of the non sample units, obtained by fitting the model to the sample data.

Linear Mixed Model

The linear mixed model (LMM) assumes that the population variables Y_{di} follow the model

$$\begin{aligned} &Y_{di} = \mathbf{x}'_{di}\boldsymbol{\gamma} + u_d + e_{di}, \\ &u_d \overset{iid}{\sim} N(0, \sigma_e^2), \quad e_{di} \overset{iid}{\sim} N(0, \sigma_e^2), \quad i = 1, \ldots, N_d, \quad d = 1, \ldots, D. \end{aligned}$$

The EBLUP under this model is given by

$$\hat{P}_d^{LMM} = \frac{1}{N_d}\left(\sum_{i \in s_d} Y_{di} + \sum_{i \in \bar{s}_d} \hat{Y}_{di}^{LMM}\right), \quad d = 1, \ldots, D,$$

where $\hat{Y}_{di}^{LMM} = \mathbf{x}'_{di}\hat{\boldsymbol{\gamma}} + \hat{u}_d$.

Appendix 2: Description of STATPOP data set

VARIABLE	DESCRIPTION	VALUE	LABEL
Person_pseudo	Pseudonym	1 - 6662333	
age	Age	≥ 15	
sex	Gender	1 2	Man Woman
maritalStatus	Marital status. 8 categories, one of which for missing values.	1 2 3 4 5 6 7 -9	Single Married Widowed Divorced Unmarried In a registered partnership Partnership dissolved No indication

nationalityid	Nationality id. 8100 = CH.	8100 to 9999 -1 -6 8000	Country/territory Stateless Not attributable Foreigner (country not indicated)
Residencepermit	Residence permit. -2 = no residence permit needed.	01 to 13 -2 -9	Foreigner category (eCH-0006) Swiss No indication
Typeofresidence	Type of residence. Only individual with a main domicile are considered.	1 2 3	Main domicile Secondary domicile No main domicile in Switzerland
secondaryResidenceId1	Secondary residence Municipality id of first secondary residence.	1 to 9999 empty	Commune Not applicable
populationtype	Type of population	1	permanent resident population
TypeofHousehold	Type of household	1	private
HouseholdSize	Size of the household		
Strata	Strata of the unit	AG00 AI00 AR00 BE00 BE02 BL00 BS00 FR00 GE00 GL00 GR00 JU00 LU00 NE00 NW00 OW00 SG00 SH00 SO00 SZ00 TG00 TI00 UR00 VD00 VS00 ZG00 ZH00	Aargau Appenzell Innerhoden Appenzell ausserrhoden Bern Bern Basel-Landschaft Basel-Stadt Fribourg Genève Glarus Graubünden Jura Luzern Neuchâtel Nidwalden Obwalden St.Gallen Schaffhausen Solothurn Schwyz Thurgau Ticino Uri Vaud Valais Zug Zürich
StrataSize	Size of the strata		
ReportingMunicipalityid	Municipality id	1 - 6810	2485 municipalities

localitysize	Municipality size		
District	District	101 to 2603	147 Districts
Canton	Canton	1-26	1=Zürich 2=Bern 3=Luzern 4=Uri 5=Schwyz 6=Obwalden 7=Nidwalden 8=Glarus 9=Zug 10=Fribourg 11=Solothurn 12=Basel-Stadt 13=Basel-Landschaft 14=Schaffhausen 15=Appenzell ausserrhoden 16=Appenzell Innerhoden 17=St.Gallen 18=Graubnden 19=Aargau 20=Thurgau 21=Ticino 22=Vaud 23=Valais 24=Neuchtel 25=Genve 26=Jura
NUTS2	NUTS2	1 2 3 4 5 6 7	22,23,25 2, 10, 11, 24, 26 12, 13, 19 1 8,14,15,16,17,18,20 3,4,5,6,7,9 21
In_OASI	0-1 if unit is/is not in OASI in 2011		
Income	Income in 2011 (truncated above 100,000 CHF)		
January	0-1 if contributed/not contributed to OASI in Jan. 2011		
February	0-1 if contributed/not contributed to OASI in Feb. 2011		
March	0-1 if contributed/not contributed to OASI in Mar. 2011		
April	0-1 if contributed/not contributed to OASI in Apr. 2011		
May	0-1 if contributed/not contributed to OASI in May. 2011		
June	0-1 if contributed/not contributed to OASI in Jun. 2011		
July	0-1 if contributed/not contributed to OASI in Jul. 2011		
August	0-1 if contributed/not contributed to OASI in Aug. 2011		
September	0-1 if contributed/not contributed to OASI in Sep. 2011		

October	0-1 if contributed/not contributed to OASI in Oct. 2011		
November	0-1 if contributed/not contributed to OASI in Nov. 2011		
December	0-1 if contributed/not contributed to OASI in Dec. 2011		
In_survey	0-1 if the unit took part to the survey or not		
Active	Variable of interest Reference date: 31.12.	0-1	not active/active
Weight	Design weight		