Supplementary material: Constructing socio-demographic indicators for National Statistical Institutes using mobile phone data: estimating literacy rates in Senegal

Timo Schmid[*], Fabian Bruckschen[*], Nicola Salvati[**], and Till Zbiranski[*]

[*]Institute of Statistics and Econometrics, Freie Universität Berlin, Berlin, Germany
[**]Economic Department, University of Pisa, Pisa, Italy

Additional description: mobile phone covariates

Table 1 describes the covariates used in the paper. The variables are split by categories to ease the understanding of their calculation and origin. Hourly covariates have been calculated on hourly call detail records, daily covariates on aggregated daily call detail records and so on. The variables in the category *interactions* take every single interaction for the year 2013 into account. The covariates are first calculated on a tower level for the year 2013 and then the median is applied for the higher geographic levels like communes and regions. For instance, the covariate *ic_sms_work_ratio* for a tower is the ratio of incoming SMS during 9am to 5pm over all incoming SMS for the year 2013 based on hourly call detail records.

Additionally to the variables described in Table 1 we created covariates with the open-source python toolkit bandicoot (http://bandicoot.mit.edu) (Montjoye et al., 2013). A list of these variables can be found here http://bandicoot.mit.edu/docs/reference/index.html.

Design-based simulation for unemployment

The results presented in Table 2 split by the 191 in-sample, the 210 out-of-sample and the 30 out-of-covariate communes. The table reports summary statistics of the RMSE and Bias of the benchmarked estimators (FH Bench, NL Bench, and NLRS Bench) over communes. The performance of the FH Bench and NLRS Bench is very comparable regarding Bias and RMSE for the in-sample and out-of-sample communes and outperforms the NL Bench estimator in this particular simulation study. For the out-of-covariate communes, where the covariates are obtained by geographically weighting as described in Section 2, all benchmarked model-based estimators (FH Bench, NL Bench, and NLRS Bench) reveal on average a small positive bias. In addition, we point out that the results of the benchmarked estimators (FH Bench, NL Bench, and NLRS Bench) are very similar to the non-benchmarked estimators (FH Trans, NL, and NLRS) because the average of the commune level estimates required only a small adjustment to meet the national estimate for the country.

Table 1: Mobile phone covariates

Name	Covariate	Description
Distance		
dist2d	distance to Dakar	The distance to the centroid of the Dakar region in kilometers.
calls_dist_mean	average calls distance	The average distance between towers that were involved in call interactions during the year in kilometers.
sms_dist_mean	average SMS distance	The average distance between towers that were involved in SMS interactions during the year in kilometers.
Interactions		
calls_entropy	entropy of calls	The entropy of calls based on tower to tower interactions throughout the whole year.
sms_entropy	entropy of SMS	The entropy of SMS based on tower to tower interactions throughout the whole year.
calls_isolation	isolation of calls	Total number of towers that a tower had call interactions with. The lower this number, the more isolated a tower is assumed to be in terms of calls.
sms_isolation	isolation of SMS	Total number of towers that a tower had SMS interactions with. The lower this number, the more isolated a tower is assumed to be in terms of SMS.
Based on yearly aggregates		
calls_ratio	calls ratio	The ratio of outgoing calls over incoming calls.
sms_ratio	SMS ratio	The ratio of outgoing SMS over incoming SMS.
vol_ratio	call volume ratio	The ratio of minutes from outgoing calls over minutes from incoming calls.
sms2calls_ratio	SMS to calls ratio	The ratio of outgoing SMS over outgoing calls.
calls2d_ratio	calls to Dakar ratio	The ratio of call interactions where a tower inside the Dakar region was involved over all call interactions.
sms2d_ratio	SMS to Dakar ratio	The ratio of SMS interactions where a tower inside the Dakar region was involved over all SMS interactions.
Based on monthly data		
calls_ratio_var	variance of calls ratios	The variance of the monthly ratios of outgoing calls over incoming calls.
sms_ratio_var	variance of sms ratios	The variance of the monthly ratios of outgoing sms over incoming sms.
vol_ratio_var	variance of call volume ratios	The variance of the monthly ratios of outgoing call minutes over incoming call minutes.
Based on daily data		
og_calls_week_ratio	outgoing calls week ratio	The percentage of calls being initiated during the weekend.
og_sms_week_ratio	outgoing SMS week ratio	The percentage of SMS being sent during the weekend.
og_vol_week_ratio	outgoing call volume week ratio	The percentage of minutes from outgoing calls during the weekend.
ic_calls_week_ratio	incoming calls week ratio	The percentage of calls being received during the weekend.
ic_sms_week_ratio	incoming SMS week ratio	The percentage of SMS being received during the weekend.
ic_vol_week_ratio	incoming call volume week ratio	The percentage of minutes from incoming calls during the weekend.
Based on hourly data		
og_calls_work_ratio	outgoing calls work ratio	The ratio of outgoing calls during 9 am to 5 pm over all outgoing calls.
og_sms_work_ratio	outgoing SMS work ratio	The ratio of outgoing SMS during 9 am to 5 pm over all outgoing SMS.
og_vol_work_ratio	outgoing call volume work ratio	The ratio of minutes from outgoing calls during 9 am to 5 pm over all outgoing minutes.
ic_calls_work_ratio	incoming calls work ratio	The ratio of incoming calls during 9 am to 5 pm over all incoming calls.
ic_sms_work_ratio	incoming SMS work ratio	The ratio of incoming SMS during 9 am to 5 pm over all incoming SMS.
ic_vol_work_ratio	incoming call volume work ratio	The ratio of minutes from incoming calls during 9 am to 5 pm over all incoming minutes.
og_calls_peak_ratio	outgoing calls peak ratio	The ratio of calls being initiated between 3 to 5 am (early peak) over calls being initiated between 10 am to 12 pm (late peak)
og_sms_peak_ratio	outgoing SMS peak ratio	The ratio of SMS being sent between 3 to 5 am (early peak) over sms being sent between 10 am to 12 pm (late peak)
og_vol_peak_ratio	outgoing call volume peak ratio	The ratio of minutes from outgoing calls between 3 to 5 am (early peak) over minutes of outgoing calls between 10 am to 12 pm (late peak)
ic_calls_peak_ratio	incoming calls peak ratio	The ratio of calls being received between 3 to 5 am (early peak) over calls being received between 10 am to 12 pm (late peak)
ic_sms_peak_ratio	incoming SMS peak ratio	The ratio of SMS being received between 3 to 5 am (early peak) over SMS being received between 10 am to 12 pm (late peak)
ic_vol_peak_ratio	incoming call volume peak ratio	The ratio of minutes from incoming calls between 3 to 5 am (early peak) over minutes of incoming calls between 10 am to 12 pm (late peak)

Table 2: Performance of benchmarked predictors over communes in design-based simulations

191 In-sample communes

Indictor	Estimator	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
RMSE	FH Bench.	0.017	0.030	0.042	0.053	0.069	0.254
	NL Bench.	0.014	0.040	0.049	0.056	0.060	0.262
	NLRS Bench.	0.015	0.029	0.043	0.053	0.070	0.256
Bias	FH Bench.	-0.196	-0.023	0.006	0.007	0.035	0.253
	NL Bench.	-0.103	-0.009	0.005	0.012	0.028	0.171
	NLRS Bench.	-0.204	-0.023	0.005	0.006	0.035	0.255

210 Out-of-sample communes

		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
RMSE	FH Bench.	0.009	0.032	0.055	0.074	0.104	0.344
	NL Bench.	0.009	0.035	0.061	0.076	0.104	0.322
	NLRS Bench.	0.008	0.031	0.056	0.073	0.103	0.344
Bias	FH Bench.	-0.343	-0.039	0.017	0.013	0.068	0.252
	NL Bench.	-0.321	-0.040	0.014	0.015	0.067	0.284
	NLRS Bench.	-0.344	-0.038	0.013	0.013	0.064	0.253

30 Out-of-covariate communes

		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
RMSE	FH Bench.	0.010	0.036	0.064	0.081	0.102	0.282
	NL Bench.	0.010	0.043	0.077	0.086	0.101	0.282
	NLRS Bench.	0.010	0.038	0.064	0.082	0.100	0.284
Bias	FH Bench.	-0.168	0.003	0.049	0.042	0.095	0.282
	NL Bench.	-0.150	-0.001	0.051	0.046	0.096	0.282
	NLRS Bench.	-0.165	0.003	0.044	0.042	0.090	0.284

References

Montjoye, Y.-A., J. Quoidbach, F. Robic, and A. S. Pentland (2013). *Social Computing, Behavioral-Cultural Modeling and Prediction: 6th International Conference, SBP 2013, Washington, DC, USA, April 2-5, 2013. Proceedings*, Chapter Predicting Personality Using Novel Mobile Phone-Based Metrics, pp. 48–55. Berlin, Heidelberg: Springer Berlin Heidelberg.