Wednesday 26 November 2025 Under review at Sport England

Small Area Estimates of Physical Activity in England

Jo Kroese
Dataknit
jo@jokroese.com

ABSTRACT This technical report describes the creation of small area estimates of physical activity in England.

KEYWORDS physical activity; small area estimates; multilevel regression and poststratification (MRP); Bayesian modeling; Bayesian workflow; survey methodology

Table of contents

Executive Summary

- This analysis provides unprecedented level of detail in small area estimates of physical activity for adults and children and young people.
- For adults,

  - there is significant variation in the level of activity between LSOAs, with percentage of adults physically active ranging between 36.0% to 82.9%.
  - this variation is mostly driven by wide disparities at an individual level, with some adults being predicted as having a 2.1% probability of being physically active, up to some who have a 90.0%.
  - the strongest predictors of physical activity are at an individual level: age, disability and the National Statistics Socio-economic classification (NS-SEC).
- For children and young people, the understanding of key predictors remains at a much earlier stage.

Introduction

Physical activity plays a key role in the health of individuals, communities and society. However, in England these benefits are not spread equally: one's likelihood of participating in physical activity strongly depends on who you are and where you live.

As much of the work to improve lives and communities happens locally, this lack of estimates is currently impeding Sport England's long-term vision.

This technical report aims to close this gap. With robust small area estimates to show which groups and areas are most in need, Sport England and their local partners can use their resources more effectively. Basing these decisions on evidence will prove crucial for Sport England to achieve their vision outlined in the new 'Uniting the Movement' strategy (Sport England 2021) to harness the power of sport and physical activity.

The technical report describes the creation of small area estimates of physical activity that are unprecedented in their precision, reaching to LSOA level and allowing demographic breakdowns within these areas. Specifically, the results are:

- Small area estimates of physically active, physically inactive and participating in sports at MSOA and LSOA level.
- It also includes breakdowns of these small area estimates by demographic variables: age, gender, ethnicity, NS SEC, age group and Sport England's inequality metric.

Methodology

The work is based on a multilevel regression and poststratification (MRP) analysis. Whilst standard design-based survey methodologies have historically struggled with small-area estimation, MRP is able to accurately reconstruct small areas. This has led to MRP becoming the gold standard in small area estimates and arguably survey statistics in general.

MRP began as a tool in political science for estimating presidential elections in the US. Indeed, implementations of the technique powered the only polls to correctly predict the Brexit referendum result and the US election, both in 2016. It has since been used in a broad range of applied problems ranging from epidemiology to social science. Our work at Substance and Dataknit has pushed it further, including the first use of MRP in marine science and using it to generate the first estimates of physical activity rates at LSOA level.

MRP combines two aspects: a multilevel model (in this case, estimating the physical activity or participation of a person) and poststratification.

As with all our work on small area estimates, our model is built with Bayesian techniques. The choice to use a Bayesian approach, and bypass a frequentist approach, is now standard for generating small area estimates. Some key reasons are that Bayesian methods:

- Provide an easy and consistent way to do complex modelling such as multilevel modelling (for partial pooling of data, supporting groups with few data points), missing data imputation (which can avoid throwing away incomplete survey data) and spatial correlation (useful for representing similarities between neighbouring areas)
- Give a rich output, obtaining full distributions for all parameters of interest, not just point estimates and standard error estimates
- Quantify all sources of uncertainty in the model.

I developed the model using the 'Bayesian Workflow', a best-practice approach to building Bayesian models [Gelman et al. (2020); @schad]. Promoted by many statisticians, including the original developers of MRP, the workflow involves a rigorous approach to building and validating models.

Whilst there are several stages to the workflow, the focus is on iterative development: building a simple model, checking the issues with it and modifying the model to fix those issues. The workflow defines tools to assess possible issues at each stage where they could come up (e.g., fake data simulation for computational issues or cross-validation for evaluating the accuracy of a model). It also suggests certain modifications that are appropriate for dealing with certain issues (e.g., adding more prior information to deal with computational issues or expanding the model if the cross-validation scores are poor). This approach, informed by decades of practice, moves the statistician in a principled way towards a robust model.

As this is a Bayesian analysis, we need priors. However, the sizes of the datasets involved mean that the priors had little impact on the results. We used weakly informative priors to support the computation process without affecting the results of the model. We checked the reasonableness of these priors through prior predictive checks.

The variables chosen are done through systematically adding 'modules' to the model, adding, for example, add in different ways. The candidate models are all compared with Leave-One-Out Cross-Validation.

It is instructive to contrast some aspects of our proposed methodology with conventional approaches used in small area estimates research. These conventional approaches have, at times, employed frequentist methods and automated variable selection methods, such as the 'forward stepwise procedure,' which have been discouraged for several decades due to limitations such as suboptimal variable selection, inflated confidence estimates, and issues related to multicollinearity in predictions. Additionally, there has been a historical practice of not thoroughly describing the model, which can impact both predictive accuracy and stakeholder understanding.

Another common issue has been the inclusion of collinear variables in the models, such as considering both individual-level age and regional covariates like 'proportion population age 65-74.' Best practices suggest that, when dealing with collinear variables, it is advisable to include only one of them, preferably the most specific one, to maintain model detail while mitigating issues associated with multicollinearity.

The principled approach of the Bayesian Workflow is designed precisely to address these issues and improve the quality of small area estimates research.

The end result is to produce four separate models. For adults, there are three binary logistic models:

- To estimate if an adult participates in sport or not
- To estimate if an adult is active or not
- To estimate if an adult is inactive or not

For children, there is one model:

- To estimate the minutes of physical activity per week of a child.

This approach combines simplicity and ease to explain to stakeholders.

All the Bayesian models were fit using brms (Bürkner 2017), an R wrapper around Stan (2023), a 'state-of-the-art platform for statistical modeling and high-performance statistical computation'. Stan is currently the most advanced software for fitting Bayesian models and is used throughout small area estimates and MRP.

The poststratification frame needed is more detailed than what the ONS releases from the Census. Therefore, we use Iterative Proportional Fitting to combine multiple marginals into a more detailed joint distribution.

To produce small area estimates from the model, we used poststratification. Post-stratifying is made easy, flexible and robust by using tidymrp (Jo Kroese 2023). Through tidymrp, we can easily turn the fitted model into estimates and uncertainty intervals for all subgroups. This allows us to provide the breakdowns for any possible demographic category.

Local Authority results are calculated both to the December 2022 boundaries and April 2023 boundaries, which merged 17 Local Authorities into 4. The estimates for the few Local Authorities that differ between them were calculated through a weighted combination of the constituent Local Authorities.

Data

The data is focused on the Active Lives dataset. The Active Lives dataset is one of the most comprehensive datasets of physical activity in the world. This analysis makes use of its full potential at a governmental level to provide highly detailed estimates of the physical activity levels of adults, children and young people.

The project also makes use of the IMD variables.

The data from all the Active Lives surveys was imported and tidied with R (Henry and Wickham 2023), using the 'tidyverse' packages (Wickham et al. 2019).

Results

Following the iterative workflow, for the adults I arrived at a logistic model that had the following independent variables:

- disability, NS SEC, ethnicity, gender as random effects
- age as a smooth term, with an interaction with disability

- local authority code as a random effect

For the children and young people, I arrived at a hurdle gamma model that had the following independent variables:

- disability, ethnicity, gender, family affluence as random effects
- age as a smooth term
- Local Authority as a random effect

To arrive at a poststratification frame with enough detail, I combined 13 marginal distributions for the adults and 15 for children and young people.

The results for adults show significant variation of activity at LSOA level throughout England (@fig-lsoa-results) with the lowest value of percentage active being 40.1% and the highest being 82.2%. This variation is driven through wide disparities at an individual level, with some individuals being predicted as having a 2.1% probability of being physically active, up to some who have a 90.0% probability.

In children and young people, individual factors play a far smaller role with much of the variation coming at a level between individual and Local Authority.

Discussion

The results show that adults' activity is relatively simple to predict: it is mostly driven by age, NS SEC and disability. Other variables can further help the model but provide limited additional support.

In children and young people, the picture is much more murky as there are no clear variables that are in Active Lives and the Census that provide strong predictors of their activity.

Limitations

One limitation is the measurement of 'activity' in adults. Due possibly to the role of intense activity, that counts for two times hours of moderate activity, some activity values are unusually high. Further investigation into the metric reveals a very unusually shaped distribution.

However, this limitation only strongly affects very high activity individuals. The model I built squashes the activity levels to 3 levels: inactive, less active and active. This reduced its dependence on the extremes of this metric.

In terms of children and young people, there a more limited understanding of what contributes to activity. This makes it more difficult to model and so the resulting estimates should be considered with more caution.

Recommendations

In terms of policy, the estimates provide a literal map of how to target resources towards low activity areas. The breakdowns by demographics provide a further focus on where interventions should be targeted.

In terms of research, the analysis raises questions around the activity of children and young people. The most fundamental is: what drives it? It is clearly very different to activity in adults and seems to be primarily driven by factors happening at the local

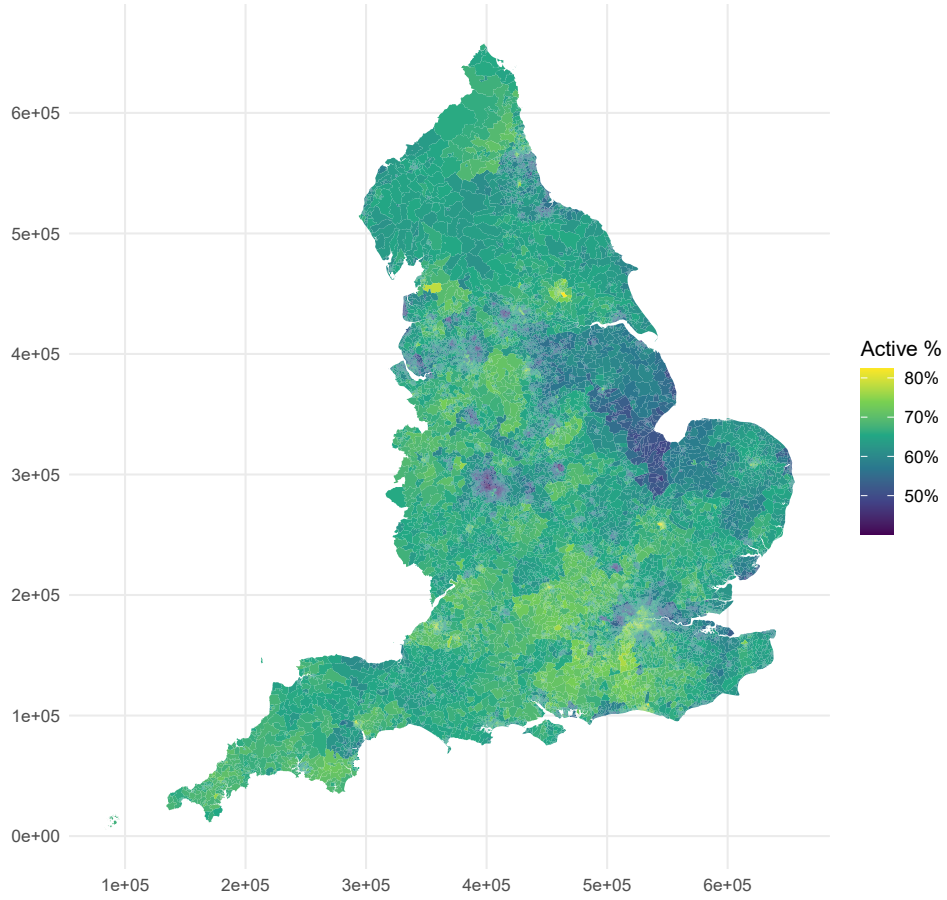

Figure 1: Chloropleth plot of percentage of adults physically active in each LSOA in England.

level. A crucial step to moving forwards with improving activity for children and young people is to understand this.

Once there are hypotheses of what these factors could be, we should consider how to capture more data relevant to these factors in the Active Lives survey.

Conclusion

This analysis provides the most precise estimates of physical activity produced in not just England but, of those released publicly, the world. Created following best practices in small area estimates, they provide a view of physical activity across England. The results can be used as a centre of Sport England's and local partners' framework to increase physical activity in England.

References

Bürkner, Paul-Christian. 2017. '{Brms}: An {r} Package for {Bayesian} Multilevel Models Using {Stan}' 80. https://doi.org/10.18637/jss.v080.i01.

Gelman, Andrew, Aki Vehtari, Daniel Simpson, Charles C. Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. 2020. 'Bayesian Workflow'. *arXiv:2011.01808 [Stat]*, November, 77. http://arxiv.org/abs/2011.01808.

Henry, Lionel, and Hadley Wickham. 2023. 'Rlang: Functions for Base Types and Core r and 'Tidyverse' Features'. https://CRAN.R-project.org/package=rlang.

Jo Kroese. 2023. *Tidymrp: Tidy Multilevel Regression and Poststratification (MRP)*. https://github.com/jokroese/tidymrp.

Sport England. 2021. 'Uniting the Movement', January. https://sportengland-production-files.s3.eu-west-2.amazonaws.com/s3fs-public/2021-02/Sport%20England%20-%20Uniting%20the%20Movement%27.pdf?VersionId=7JxbS7dw40CN0g21_dL4VM3F4P1YJ5RW.

Stan Development Team. 2023. '{RStan}: The {r} Interface to {Stan}'. https://mc-stan.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. 'Welcome to the {Tidyverse}' 4: 1686. https://doi.org/10.21105/joss.01686.

Appendix: Model outputs

Table 1: Summary of Fitted Model Parameters of Active Adults Model

Parameter	Estimate	SE	Lower 95% CI	Upper 95% CI
b_Intercept	-0.157	0.399	-0.954	0.655
bs_sage:disabilitydisabled_1	-3.102	3.246	-10.649	1.043
bs_sage:disabilitymissing_1	0.199	1.279	-2.214	3.059
bs_sage:disabilitynotdisabled_1	-0.853	1.595	-5.030	1.623
sd_disability__Intercept	0.577	0.300	0.210	1.351
sd_ethnicity__Intercept	0.351	0.124	0.200	0.655
sd_gender__Intercept	0.287	0.203	0.085	0.862
sd_ns_sec__Intercept	0.367	0.112	0.220	0.655
sds_sagedisabilitydisabled_1	2.283	0.768	1.184	4.077
sds_sagedisabilitymissing_1	2.747	0.941	1.314	4.990
sds_sagedisabilitynotdisabled_1	2.261	0.605	1.351	3.692
Intercept	-0.157	0.399	-0.954	0.655
r_disability[disabled,Intercept]	-0.379	0.326	-1.093	0.295
r_disability[missing,Intercept]	-0.043	0.326	-0.761	0.630
r_disability[not.disabled,Intercept]	0.370	0.326	-0.345	1.037
r_ethnicity[asian.(excl..chinese),Intercept]	-0.292	0.132	-0.560	-0.040
r_ethnicity[black,Intercept]	-0.249	0.134	-0.522	0.017
r_ethnicity[chinese,Intercept]	-0.280	0.140	-0.558	0.001
r_ethnicity[missing,Intercept]	-0.164	0.133	-0.432	0.101
r_ethnicity[mixed,Intercept]	0.336	0.137	0.056	0.603
r_ethnicity[other.ethnic.group,Intercept]	0.000	0.139	-0.289	0.274
r_ethnicity[white.british,Intercept]	0.383	0.131	0.112	0.639
r_ethnicity[white.other,Intercept]	0.264	0.132	-0.006	0.522
r_gender[female,Intercept]	0.023	0.174	-0.331	0.395
r_gender[male,Intercept]	0.198	0.174	-0.155	0.575
r_gender[missing,Intercept]	-0.141	0.216	-0.624	0.242
r_gender[other,Intercept]	-0.101	0.178	-0.478	0.264
r_ns_sec[missing,Intercept]	0.115	0.129	-0.142	0.379
r_ns_sec[ns.sec.1-2:.managerial,.administrative.and.professional.occupations,Intercept]	0.403	0.124	0.155	0.662
r_ns_sec[ns.sec.3:.intermediate.occupations,Intercept]	0.022	0.124	-0.230	0.280
r_ns_sec[ns.sec.4:.self.employed.and.small.employers,Intercept]	0.017	0.125	-0.234	0.274
r_ns_sec[ns.sec.5:.lower.supervisory.and.technical.occupations,Intercept]	-0.066	0.125	-0.314	0.190
r_ns_sec[ns.sec.6-7:.semi-routine.and.routine.occupations,Intercept]	-0.286	0.124	-0.536	-0.029

r_ns_sec[ns.sec.8:.long.term.unemployed.or.never.worked,Intercept]	-0.556	0.126	-0.807	-0.301
r_ns_sec[ns.sec.9:.other./.unclassified,Intercept]	-0.190	0.126	-0.444	0.068
r_ns_sec[ns.sec.9:.students,Intercept]	0.436	0.129	0.185	0.695
s_sagedisabilitydisabled_1[1]	-2.635	1.631	-6.031	0.404
s_sagedisabilitydisabled_1[2]	1.030	1.556	-2.251	3.631
s_sagedisabilitydisabled_1[3]	1.255	0.640	0.244	2.753
s_sagedisabilitydisabled_1[4]	1.625	0.576	0.385	2.627
s_sagedisabilitydisabled_1[5]	0.531	0.984	-1.257	2.625
s_sagedisabilitydisabled_1[6]	4.211	1.105	1.857	5.999
s_sagedisabilitydisabled_1[7]	0.301	1.163	-2.061	2.517
s_sagedisabilitydisabled_1[8]	-1.796	1.481	-4.778	0.901
s_sagedisabilitymissing_1[1]	0.400	2.187	-3.658	5.090
s_sagedisabilitymissing_1[2]	3.117	1.136	0.983	5.392
s_sagedisabilitymissing_1[3]	-0.075	0.451	-0.976	0.782
s_sagedisabilitymissing_1[4]	3.962	0.657	2.732	5.303
s_sagedisabilitymissing_1[5]	-0.782	1.230	-3.187	1.605
s_sagedisabilitymissing_1[6]	2.329	1.103	0.280	4.599
s_sagedisabilitymissing_1[7]	1.482	1.829	-1.823	5.435
s_sagedisabilitymissing_1[8]	4.247	2.053	0.669	8.549
s_sagedisabilitynotdisabled_1[1]	-3.111	1.080	-5.272	-1.024
s_sagedisabilitynotdisabled_1[2]	2.464	0.849	0.431	3.961
s_sagedisabilitynotdisabled_1[3]	1.231	0.363	0.600	2.068
s_sagedisabilitynotdisabled_1[4]	3.475	0.349	2.711	4.093
s_sagedisabilitynotdisabled_1[5]	0.096	0.719	-1.251	1.542
s_sagedisabilitynotdisabled_1[6]	2.789	0.616	1.366	3.866
s_sagedisabilitynotdisabled_1[7]	0.836	0.812	-0.785	2.403
s_sagedisabilitynotdisabled_1[8]	-0.203	1.027	-2.299	1.754
lprior	-17.676	2.767	-23.633	-13.057
lp__	-104569.163	18.747	-104606.000	-104534.000