Friday, 24 November 2023 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, Betancourt, and Vasishth, n.d.). 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. Poststratifying 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 (Figure 1) with the lowest value of percentage active being 36.0% and the highest being 82.9%. 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.

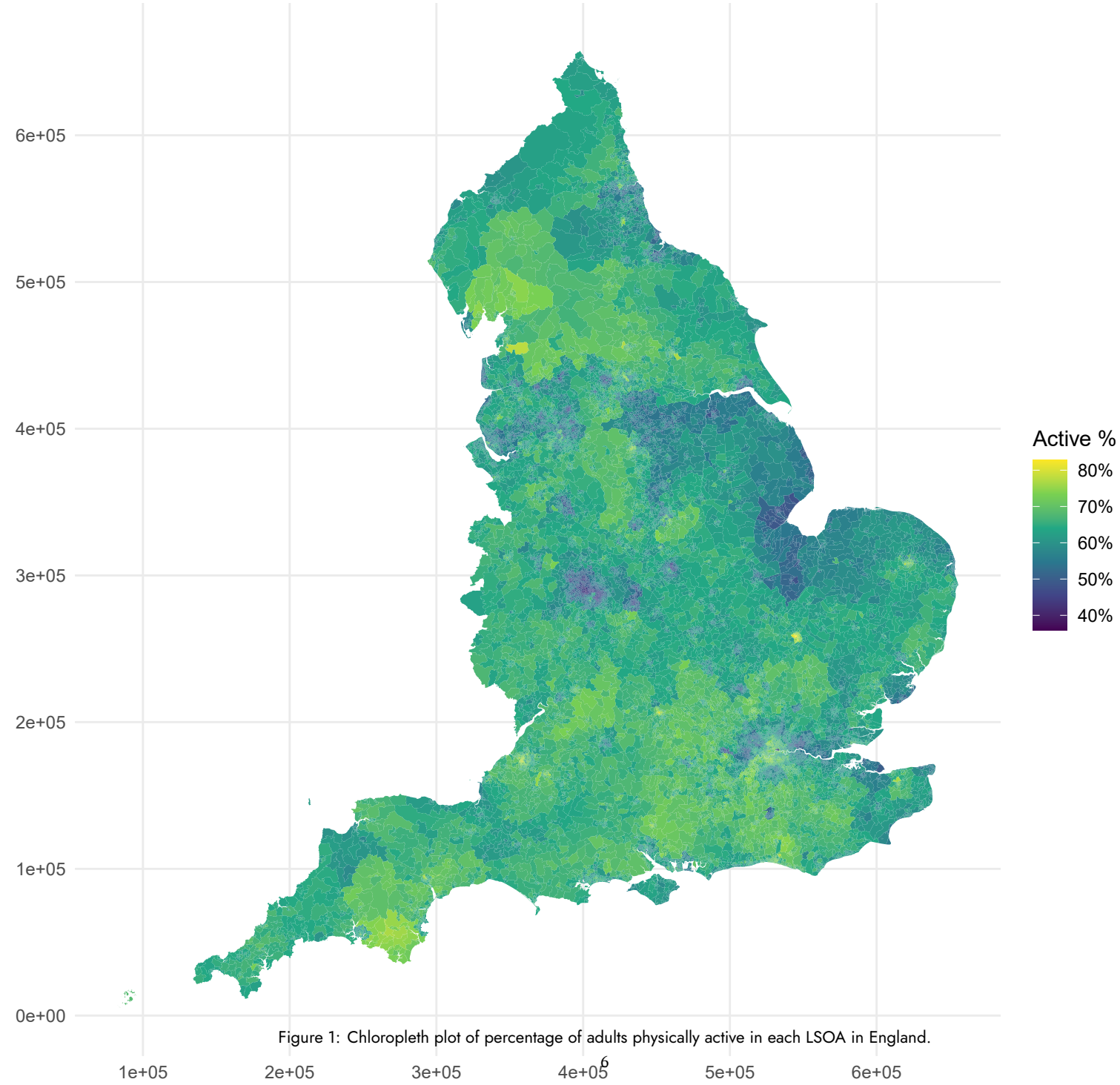

Figure 1: Chloropleth plot of percentage of adults physically active in each LSOA in England.

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 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.

Schad, Daniel J., Michael Betancourt, and Shravan Vasishth. n.d. 'Toward a Principled Bayesian Workflow in Cognitive Science'. https://doi.org/10.48550/arXiv.1904.12765.

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=7JxbS7dw40CNog21_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.

Tuesday, 31 October 2023 Under review at Sport England

Frequently Asked Questions

Jo Kroese
Dataknit
jo@jokroese.com

ABSTRACT Frequently Asked Questions on the statistical analysis of physical activity in England.

KEYWORDS physical activity; public health; statistical modelling; local authorities; demographic analysis

How can you estimate values for populations which are very small?

The analysis gives estimates for all possible cross-sections of the variables included in the model. For every possible cross-section, the model gives an estimates of the expected physical activity of someone from this group. This includes large populations — such as white, 35-year-old, university-educated men in a specific area of London — and also very small populations — such as Chinese, 99-year-old, female students in a rural, mostly white area.

These small populations will have larger uncertainties but are still given in the estimates. Their effect on any of the estimates is proportional the size of the strata (i.e. the number of people with that description).

Why aren't there estimates for certain cross-sections?

The poststratification frame — the data on how many people of each cross-section there are — is built using Iterative Proportional Fitting. This technique is able to bring together various datasets from the Office of National Statistics to create a more detailed frame than they release whilst aligning with all of their released data.

For some cross-sections, there are no people fitting the cross-section's description. While the model is able to give estimates for these fictional people, we omitted them in the released data.

Why aren't the published Active Lives results aligned completely with the small area estimates?

Sport England publish Local Authority estimates of physical activity from Active Lives. hese estimates are based on survey weighting. he small area estimates are calculated using a different technique — multilevel regression and poststratification — which is especially proficient at providing estimates for areas with limited data, such as small area estimates.

he two different calculation methods lead to small variations in the estimates at Local Authority level.

Why isn't there as much variation at an LSOA/MSOA level in the CYP data?

he CYP estimates include variation due to different demographics (age, disability, family affluence and ethnicity) at an LSOA level. However, these variables are of much less importance for CYP than they are for adults.

Local Authority has an impact and is included as a key driver in the model. However, for data privacy reasons, the dataset does not include any information that ties the respondent to any closer geography. his leads to difficulties in representing the likely variation at LSOA level of physical activity.

he closest we can get is the Index of Multiple Deprivation (IMD) value which is specific to LSOAs. While this could be important, the IMD value in the dataset is of the school, not of the child's LSOA of residence. his removes some of its effectiveness for mapping the effect of deprivation on physical activity. Whilst it is used in the model, it is not able to capture the true impact that deprivation has on physical activity. his in turn creates lower variation in the small area estimates.

What the main drivers of physical activity rates in small areas?

For adults, the main drivers are individual level: disability status, age and National Statistics Socio-economic Classification (NS SEC).

For children, the drivers appear to be concentrated at a geographic level lower than Local Authority. For individual level predictors, disability and age have significantly less impact in children and whilst family affluence — a rough analogue to NS SEC — has an impact, it is not as large. here is evidence that a significant driver of children's physical activity is roughly at school level. However, the available data on this means we are still exploring the key drivers.