Methodology overview for Subregional Pupil Projections

December 2024

Table of Contents

1. Introduction

The Greater London Authority is launching a new series of Subregional Pupil Projections, intended to serve as a resource to help inform policy, service delivery, and strategic planning work. This document provides an overview of the rationale for this new output, together with the methodology used to produce the projections and an analysis of its accuracy.

Pupil projections are routinely used in local education planning, with detailed projections informing decisions about changes in capacity. Local authorities have a statutory requirement to provide pupil forecasts to the Department for Education (DfE) as part of the annual School Capacity survey (SCAP).

Projections commonly used in local education planning include those:
- produced in-house by the local authority
- provided to the local authority by DfE (using a model developed by ONS)
- delivered to subscribers of the GLA's School Roll Projection service
- commissioned from private sector consultants

These projections tend to be very detailed – produced at the level of Education Planning Areas or individual schools – but provide users with limited information about trends in neighbouring areas. Understanding the wider context of demand for places can be important for planners in London due to high levels of cross-border mobility, particularly among secondary pupils. Additionally, the limited public availability of such projections can make it difficult to build a coherent picture of trends in neighbouring areas[1].

[1] While projections submitted as part of the SCAP collection are subsequently published by DfE, the methods and assumptions used in each area are not transparent or consistent between areas.

Projections of future pupil numbers and demand for places are also relevant to the development and delivery of policies and programmes such as London's Universal Free School Meals scheme – as well as in spatial development strategies such as the London Plan.

These applications do not typically require projections with a high level of spatial detail, but they do place more emphasis on transparency, reproducibility, consistency, and quantification of uncertainty.

Available projections that provide a consistent view of pupil trends beyond a single authority include:

- The DfE's National Pupil Projections – underpinned by the ONS National Population Projections and available only for England as a whole.
- Pan-London Pupil Projections – most recently published by the GLA in 2018. These included consistent outputs for all electoral wards in London. However, complex production and reliance on non-publicly available input data made it too resource-intensive to regularly update these outputs.

This new output is intended to provide a consistent, public view of pupil trends across England that addresses the shortcomings in the existing outputs listed above. The features of this new set of projections are listed below.

- National Curriculum Years from Reception to Year 11 will be covered, for ten years from the last point of real data for these quantities.
- For now, we aim to project places for State-funded schools only, and to explore projections for independent schools for future projection rounds.
- It is desirable to provide uncertainty measures around projections so that users can plan within a realistic range of possible future outcomes, but few of the outputs above provide statistically-derived measures of uncertainty. We will use bootstrapping to calculate 95% prediction intervals.
- One requirement to achieve the aim above is to use a geography large enough to allow robust statistical modelling, while still producing good geographical detail below a national level. ITL Level 2 geographies (there are 33 in England in total, and 5 in London) should provide a geography level that will meet these requirements.
- This output is intended to be simple, fully transparent, replicable, and easily updateable on an annual basis. The input data sources need to be simple and reliably updated. Only two input data sources will be used, both of which are publicly available – annual live births, and pupil numbers from the school census. The GLA's population projections (which in turn depend on outputs from the ONS) won't be used as the source of births, or at any stage throughout the process.

2. Methodology

2.1. Input data sources

There are two inputs to these projections. The first is annual estimates of live births. Official estimates are published by the ONS for periods ending mid-year and are considered accurate. The input data can be found on this page for another output produced by the GLA Demography Team, as the data is already aggregated to ITL 2 level and is in an easy-to-use format.

The second data source is pupil numbers by National Curriculum year. This is released by the Department for Education and can be found here. The source for this dataset is the School Census taken annually and gives the number of pupils on roll in each National Curriculum year in January of the academic year. The measure used for number of pupils is headcount, where each student is counted as one student whether they are full-time or part-time students. Pupil counts are on the basis of location of school attended, rather than location of residence. As noted above, only state-funded schools are included.

2.2. Methodology for central forecast of projections

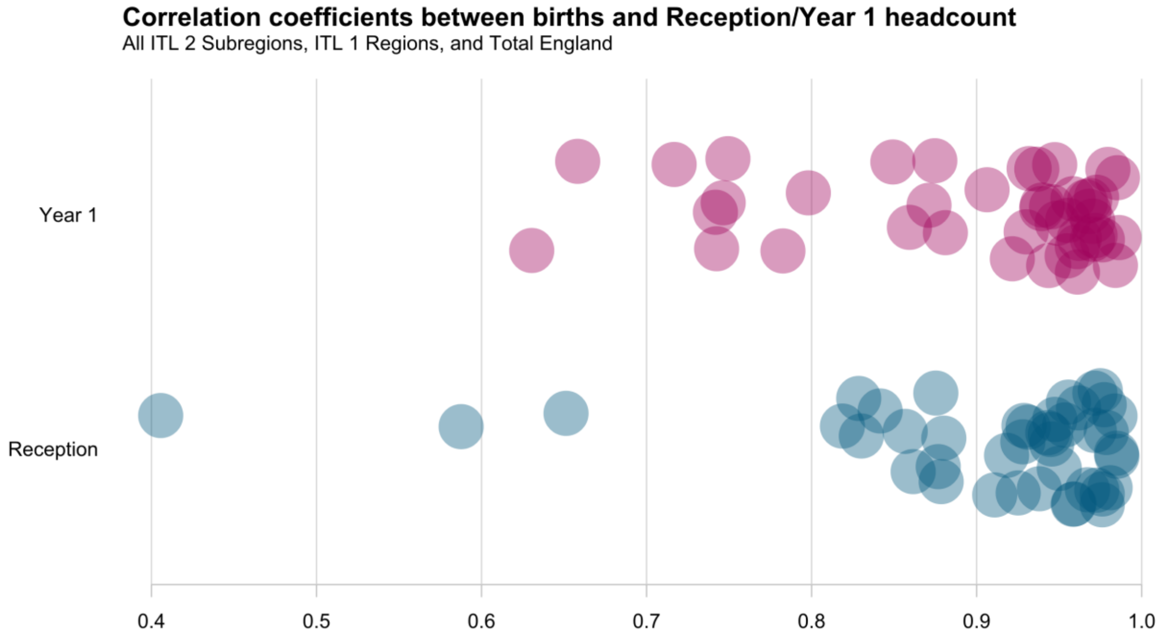
$$headcount = local\ population \times proportion\ state\ funded + net\ cross\ border$$

The key quantity of interest, pupil headcount in state-funded schools, can be expressed in the equation above. A local population of children of the appropriate age will reside in a particular geographical area, and a proportion of these will attend school in this area. There will also be a net cross border flow affecting headcount – some pupils residing in the area will attend school elsewhere, and some pupils residing elsewhere will attend school in this area. This net cross border flow is difficult to model, can vary greatly over time, and at small geographical areas can be large relative to local population.

Further, local population itself for the appropriate school-going age (e.g. 4 year olds attending Reception) is linked to births according to the equation below. A number of children will be born in an area, there will be a small and probably negligible number of deaths, and there will be net migration of this cohort from births to Reception age. Again, at smaller geographical areas net migration is often large relative to births.

$$local\ population_{cohort,t} = births_{cohort} + net\ migration_{cohort,birth \to t} - deaths_{cohort,birth \to t}$$

Therefore, to use births to project pupil numbers we need to use higher-level geographies as the relative sizes of the contribution of cross-border school attendance and net migration of 0–4-year-olds will be smaller. As illustrated in the graph below, at high geographies there is usually a close and stable relationship between births and Reception intake four years later, and for Year 1 intake five years later. Most of the correlation coefficients are above 0.9, and almost all are above 0.7. There is one subregion (Outer London - East and North East) with a particularly low correlation coefficient for births to Reception, just above 0.4. It may happen that results are worse for this geography. It is unclear at present why this number is so low, but this is something to be investigated before the next update of the projections.



However, we conclude that we can still use births to predict Reception and Year 1 numbers accurately at high geographies. The very simple formulas below are used for this purpose. The ratio is the historical ratio of births in a geography and Reception/Year 1 headcount four/five years later.

$$reception\_headcount_t = ratio\_births\_reception \times births_{t-4}$$

$$year\_one\_headcount_t = ratio\_year\_one\_births \times births_{t-5}$$

For the current set of projections, we have an actual ratio series from 2011 to 2023. To predict reception for, for example, 2025, we project forwards this ratio time series to 2025 and multiply it by births in 2021. The ratio forecast is produced with exponential smoothing, using the forecast package in R[2]. The exponential smoothing model used was Holt's trend method with multiplicative errors (there was no need to attempt to account for seasonality as the series is annual). The trend was dampened with a phi parameter set to 0.85. This dampening is relatively strong because the undampened (or minimally dampened) trend was over-projected, and produced a few unrealistically high or low ratio estimates. We also needed to project births, as the actual births series can only be used to project up to 2027 (for the 2023 run). Again we use Holt's trend method with multiplicative errors. We dampen the trend but do not set the dampening parameter, meaning it will be estimated automatically by the forecast.ets() function.

[2] The forecasting method of exponential smoothing, with varying parameters, is used for all instances in this project of casting forward a historical time series. We also dampen the trend in all instances – exponential smoothing with a damped trend is a common default forecasting method that produces good results, especially when automatic forecasts are needed. It is recommended in various guides, including Forecasting: Principles and Practice. We do not set the alpha and beta parameters – the forecast.ets() function in R will do this automatically to best fit the input time series.

For Year 2 up to Year 11, births are not used to directly calculate pupil headcount. This is because, as the National Curriculum years progress further away in time from births, the link between births and headcount becomes less reliable. Instead, these years are predicted by simulating the progression of a cohort through the National Curriculum years. A cohort in one geography will progress from Year 1 in 2021, for example, to Year 2 in 2022 with a *carryover ratio* usually close to one. It may be slightly below one (0.98 for example) or slightly above one (1.025 for example), but importantly this carryover ratio is relatively stable over time. This Year 1:Year 2 carryover ratio is modelled and is used to project the Year 2 pupil count from the Year 1 pupil that has already been projected. The same process is applied for Year 2:Year 3, Year 3:Year 4, and so on up to Year 11. Again, these ratios are modelled and projected using Holt trend method with multiplicative errors and a dampening parameter set to 0.85.

The schematic below shows how this was applied to produce Year 2 to Year 11 projections for the 2024 to 2033 set of predictions. The starting cohorts are the cells in black, and the series of carryover ratios are applied to create the projections that flow from the starting cohorts on the white/blue/green diagonals. The cells have a letter corresponding to the type of projection.

- r means there is actual data for this National Curriculum/calendar year combination.
- y means that the cell is projected from births.
- z means that the cell is projected via cohort simulation, with actual data as the starting cohort.
- x means that the cell is projected via cohort simulation, with projected year 1 data as the starting point.

	23	24	25	26	27	28	29	30	31	32	33
reception	r	y	y	y	y	y	y	y	y	y	y
yr1	r	y	y	y	y	y	y	y	y	y	y
yr2	r	z	x	x	x	x	x	x	x	x	x
yr3	r	z	z	x	x	x	x	x	x	x	x
yr4	r	z	z	z	x	x	x	x	x	x	x
yr5	r	z	z	z	z	x	x	x	x	x	x
yr6	r	z	z	z	z	z	x	x	x	x	x
yr7	r	z	z	z	z	z	z	x	x	x	x
yr8	r	z	z	z	z	z	z	z	x	x	x
yr9	r	z	z	z	z	z	z	z	z	x	x
yr10	r	z	z	z	z	z	z	z	z	z	x
yr11	r	z	z	z	z	z	z	z	z	z	z

Once this process is completed, we have a full set of projections for the time period in question. The methodology as explained in this section is applied to each geographical area separately, without accounting for possible interactions between them or constraining to national figures.

2.3. Methodology for prediction intervals around the central forecast

Uncertainty estimates are calculated for the set of projections as produced above using bootstrapping. The overall idea of bootstrapping is to create a full distribution around an estimate by resampling from the input data *n* number of times and recalculating the same estimate from the *n* number of resamples. The result is then *n* versions of the estimate of interest, distributed around the central estimate. More detail can be found on the general principle behind bootstrapping on [this blog](#).

Because the order of a time series is one of its defining characteristics, we can't randomly resample it to create the *n* resamples. Instead, we use a slightly different version of the bootstrap which involves resampling the residuals[3]. This is described in steps below, for the Reception and Year 1 forecasts that are produced from births.

1. A model is fitted to a time series. In this case, the exponential smoothing model is first fitted to births from 2011 to 2023, for a single geographical area.
2. This model will produce a fitted time series for the same years as the input series. The difference between fitted data points and actual data points is taken – these are the model residuals.
3. These residuals are then resampled *n* times. For this project, *n* is always 1,000. These 1,000 resamples are then added to the original births time series, leading to 1,000 resampled versions of the original time series. *n* is set to 1,000 as it should provide enough samples to be robust and provide the same intervals on different runs, while also avoiding unnecessary computational strain.
4. Each of these resampled time series is then forecasted using the same exponential smoothing model, from 2024 to 2033. This produces 1,000 different estimates of the births projection. These are the final bootstrapped estimates for this series.
5. The same steps above are carried out for either the births-to-reception ratio or the birth-to-Year 1 ratio. We now have 1,000 bootstrapped versions of this ratio.
6. We multiply the 1,000 bootstrapped births projection by the 1,000 versions of the ratio projection, with an appropriate lag (4 years for births to reception, 5 years for births to Year 1).
7. We now have 1,000 bootstrapped versions of the Reception projection from 2024 to 2033, because we have recreated the calculation of this projection 1,000 times using 1,000 bootstrapped samples of the inputs. We take the quantiles to match the confidence level of the prediction intervals we want. For 95% prediction intervals, this is the 2.5^{th} and the 97.5^{th} quantiles.
8. Finally, we shift these intervals such that they are symmetric around the central forecast.

For Year 2 to Year 11 we need a different bootstrapping scheme because the projections are estimated using a different method, as explained above – we start with the Year 1 cohort and simulate how this cohort will progress over time through to Year 11.

1. We already have 1,000 bootstrapped estimates of the Year 1 estimates, from the bootstrapping process above. Each of these will be used as a starting point to create a different 1,000 simulations of the cohort progression through the years.
2. We also need to reflect the uncertainty used in the carryover ratio (which is explained in section 2.2). Again to do this we use bootstrapping, by fitting the exponential smoothing model, resampling the residuals to create 1,000 samples of the original series, and forecasting each of these resamples using the same exponential smoothing model as the central forecast.
3. Then, we multiply the 1,000 simulated starting cohorts in Year 1 (from step 1) by the 1,000 simulated Year 1-Year 2 carryover ratio (from step 2) to gain 1,000 simulated versions of the Year 2 figure. After that, we multiply these 1,000 versions of Year 2 by the 1,000 simulated versions of the Year 2-Year 3 carryover ratio (again from step 2) to get 1,000 versions of Year 3. We carry on in this manner until we have 1,000 bootstrapped series of the cohort's progression from Year 2 to Year 11.

[3] A useful general introduction to residual bootstrapping for time series can be found here – Forecasting: Principles and Practice

4. Finally, we again take the 97.5th and the 2.5th quantiles of these series to get the final 95% prediction intervals around these estimates. We shift them so that they are symmetric around the central forecast.

3. Assessing performance of the projections

In this section we assess the performance of the projections against actual data (using only Reception figures). As a benchmark, we will also assess the performance of a more "naïve" method, closer to what is currently common practice in much of demographic forecasting, against actual data. We can therefore see what advantages this new method may have over simpler, alternative methods.

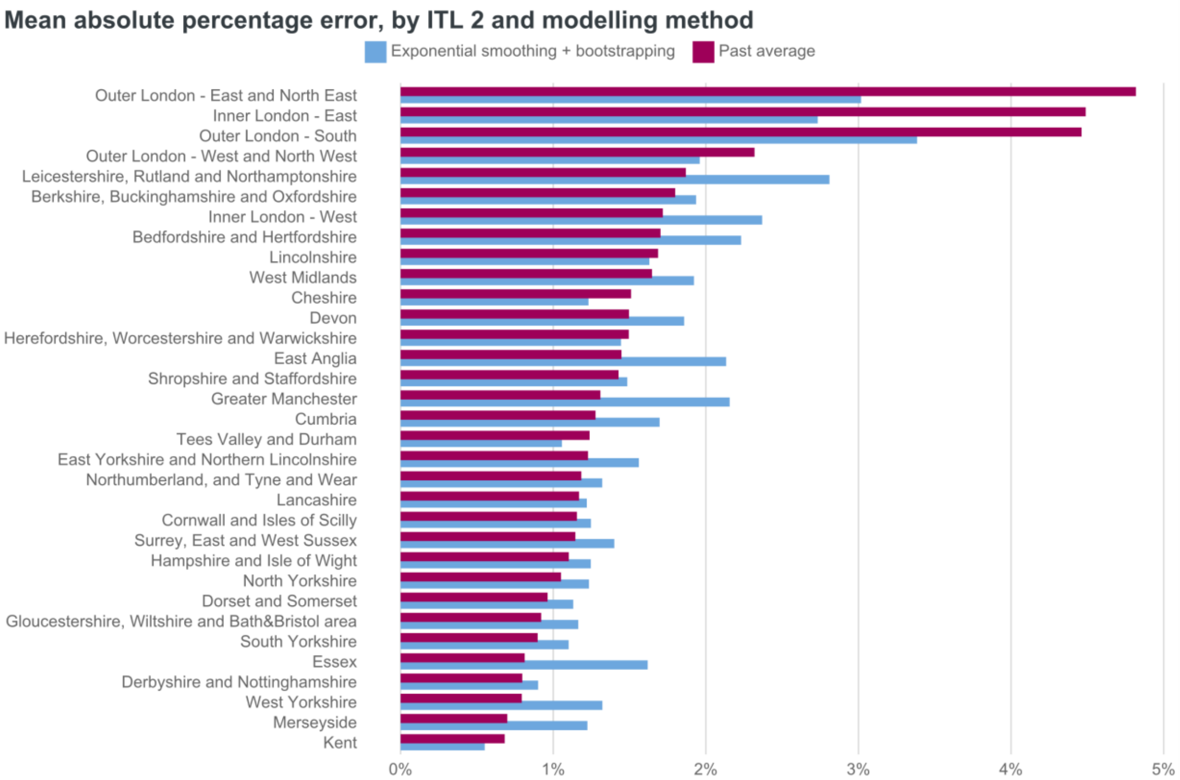
This naïve method will be referred to as the "past average" method. We use the same projected births series as before, but a different births-to-reception ratio. For the past average method, we don't use exponential smoothing to project forward the ratio series, but instead simply carry forward the average of the past *x* points in the time series. To provide a rough measure of uncertainty using this method we vary *x* to take the past 1, 3, and 5 years of past data and produce three different projections. This is not a statistical method of measuring uncertainty as we have done with bootstrapping above, so the two are somewhat incomparable. However, this is often taken by users as a quantification as uncertainty, so we will compare the performance of the two methods in accounting for uncertainty.

We needed to produce as many projected vs actual points as possible. Given the constraints of (1) having actual data from 2011 to 2022 and (2) needing at least 5 years of actual data as input for a projection, we produced projections up to 2021 using as many series of input data as possible. In total we ended up with 94 projected vs actual points for each ITL2 sub-region. The details of all possible input years can be seen in the GitHub repository here.

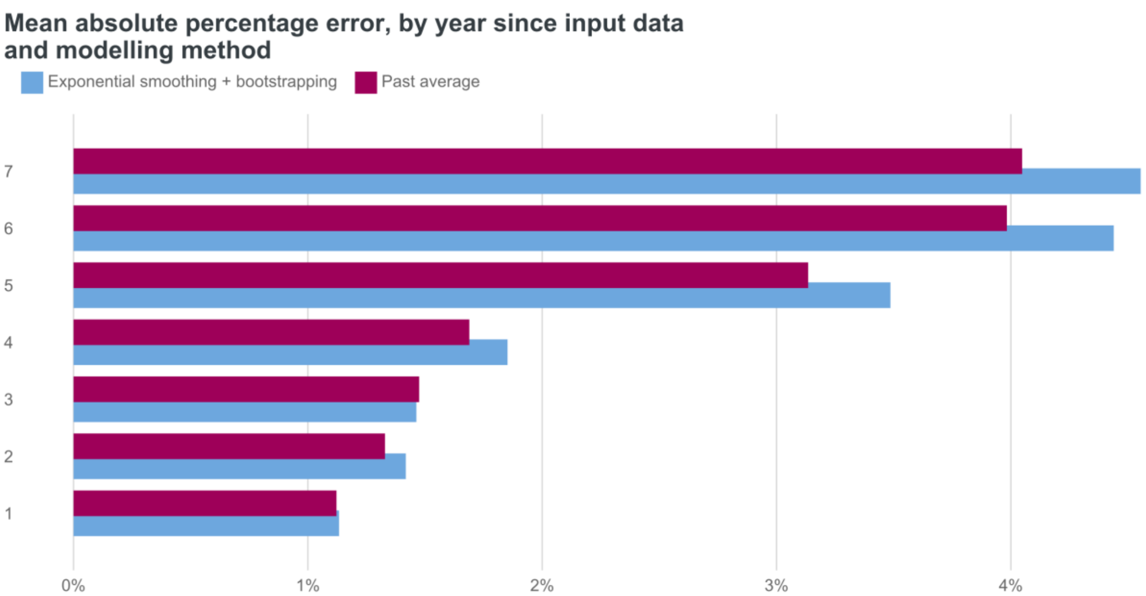
We used two main metrics to assess the performance of the projections. The first was a simple measure of accuracy called Mean Absolute Percentage Error (MAPE). The formula for this is below, where *n* is the number of observations.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|actual_i - projected_i|}{actual_i}$$

The MAPE results are shown below for each ITL2 region, for both the exponential smoothing and past average methods of forecasting. The main takeaway from this is that both methods are quite accurate, with errors ranging from 0.6% to 4.8%, without any meaningful difference in performance between the two methods.



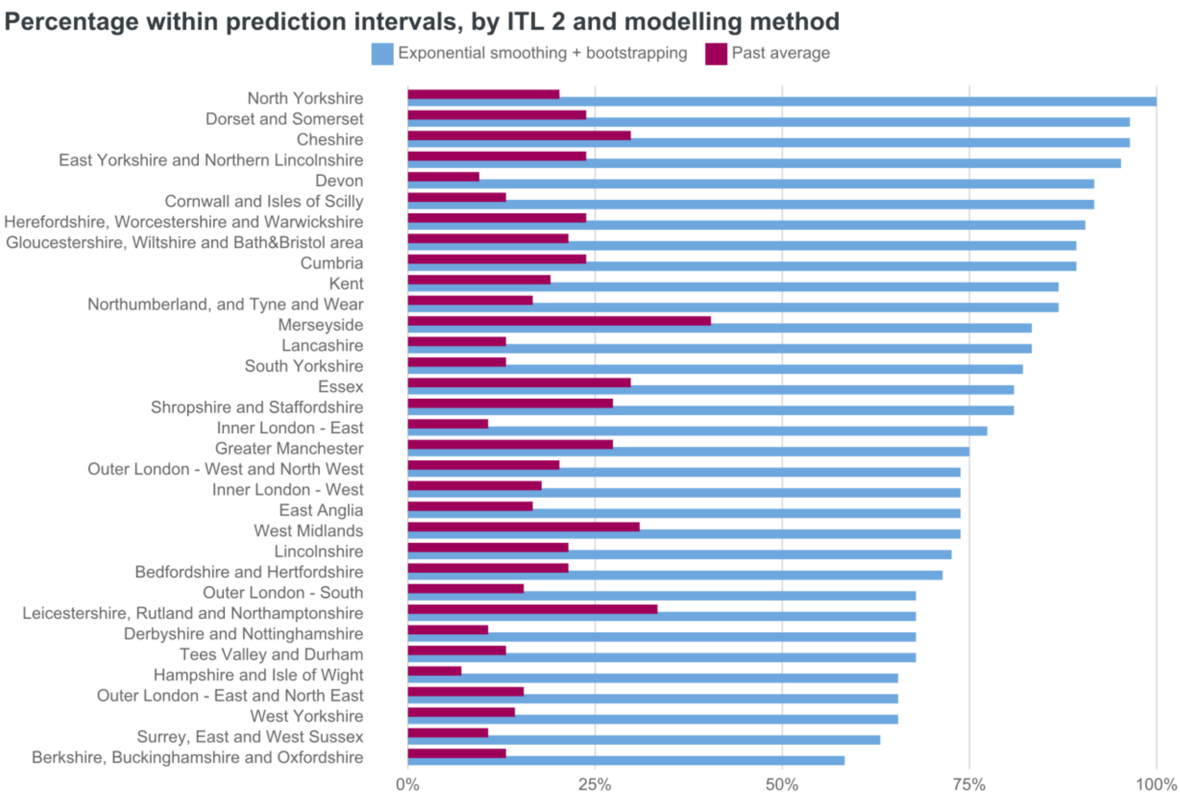
Below we again show the MAPE results, but this time broken down by the number of years since the last year of actual data (while pooling the results for all geographies together). For example, if the input data used to project was 2012 to 2016, then the 2018 projection is 2 years since the last actual data point, 2020 would be 4 years, etc. As expected, the further the projections are from the input data the less accurate they are, from just over a 1% error 1 year out to over a 4% error 7 years out. Again, it seems that there is little difference in accuracy between the two methods. Past 5 years out from actual data it does seem that the percentage errors are slightly larger for the new exponential smoothing model, but we judge the difference to be negligible.



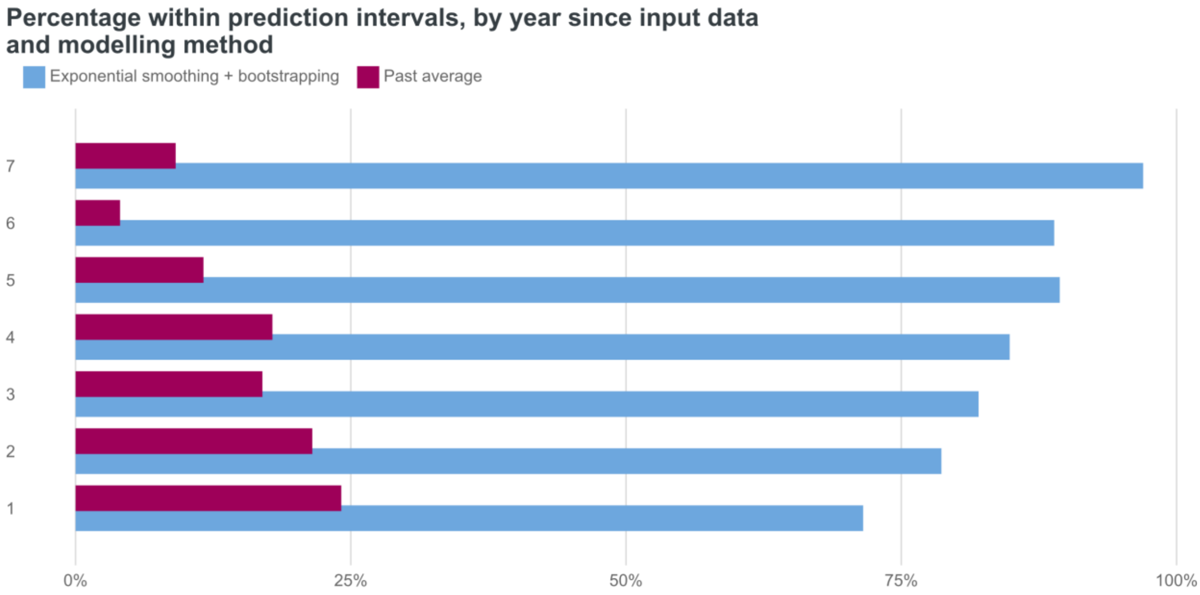
The second metric used was the percentage of actual points falling within the prediction intervals of the projected points. It is important to note that this is measuring a very different thing from the tests above – it is assessing the performance of the prediction intervals and not the accuracy of the central forecast. As a rough guide to interpreting the results, the more actual points that fall within the prediction intervals, the better. This means that users of a projection can be more confident that the range presented will capture the quantity of interest in the future.

Again, we compare the exponential smoothing with bootstrapping method against the simpler past average method. As a reminder, the past average method provides a rough estimate of uncertainty by varying how many historical points are used to take the average. Three projections are produced based on 1, 3, and 5 years of past ratio data. The highest of these three is taken as the upper bound, the lowest is taken as the lower bound, and the middle is taken as the central forecast.

We can see the results below broken down by ITL 2 sub-region for each method. Here it is immediately apparent that the new method of quantifying uncertainty outperforms the alternative. The past average method of quantifying uncertainty rarely contains more than 25% of the actual points, while the bootstrapping method always contains more than 50% of the actual points, usually at least 70%, and with an average of 79%. As they are 95% prediction intervals they should ideally contain about 95% of the actual points, but this is still a good performance. The practical significance of this is that when projections are released, users can be reasonably confident that future points will fall within the intervals provided.



Below we show the results by number of years since actual data and with results from all geographies pooled, as we have done for MAPE. Interestingly, as the projections get further out from the input data the percentage of actual points falling within the intervals increases, from around 70% for 1 year out to 97% for 7 years out. This partly reflects the narrow prediction intervals at the start of the series followed by the widening of the intervals as the years progressed. We expect to see this pattern, but it seems that they are too narrow at the start and too wide at the end of the projection series. It is unclear why the methodology followed leads to this result.



4. Limitations and next steps

Below a few limitations of the projections methodology are outlined, along with planned next steps to improve the methodology.

- As already noted, births is usually closely correlated with Reception and Year 1 headcount years later, but this is not always the case. Those geographies with a low correlation coefficient and therefore a less close relationship may have less accurate results or wider prediction intervals.
- We have only tested accuracy on Reception figures. Given that results were reasonably good for Reception we can guess that the model will perform well for Year 1 and subsequent Years, but this needs to be tested.
- This method doesn't explicitly take into account variables that we know will affect the population of young children, such as domestic and international migration. We aim to look at modelling methods that may be able to account for this.
- The forecasting method of exponential smoothing with dampening is known in the literature as an approach that usually produces good results, and in this case seems to have produced good results. However, we haven't formally tested this approach with other methods of time series forecasting such as ARIMA modelling or dynamic regression modelling. We aim to do this by assessing the projected results of several methods against actual results.
- For the exponential smoothing model, the optimisation of the alpha and beta parameters was carried out by assessing the fit between the actual series and the fitted series for the *same* input years, and the phi parameter was simply imposed as a uniform standard. It would be better to optimise all of these parameters by scanning over each of them and choosing the combination that gives the most accurate projection as compared against actual data.
- At present we haven't been able to calculate prediction intervals for Total Primary School figures and Total Secondary School Figures. We've only been able to calculate uncertainty for individual National Curriculum Years.