

October 2026

Improving the Accuracy of Housing Supply Data from Planning Applications

Edoardo Badii, Nikhil Datta and Amrita Kulka, University of Warwick

This briefing is the first in a planned series using our newly developed Housing Pipeline dataset to examine the planning system and housing supply.

Future analysis will explore how planning applications progress through the system, where delays occur, how schemes change over time, and the relationship between planning permissions and completed homes.

Summary

Governments increasingly use planning application data to assess housing delivery, monitor housing pipelines and evaluate planning reforms. However, counting proposed homes is harder than it appears. Planning portals are administrative portals, not research datasets. Records are drawn from hundreds of local authorities, with labelling conventions that vary across authorities, over time and even between staff entering records.

A new dataset, the Housing Pipeline, offers a more reliable picture. It draws on approximately 19 million planning applications across 400 authorities and is derived from processing over 120 terabytes of data.

The analysis finds that:

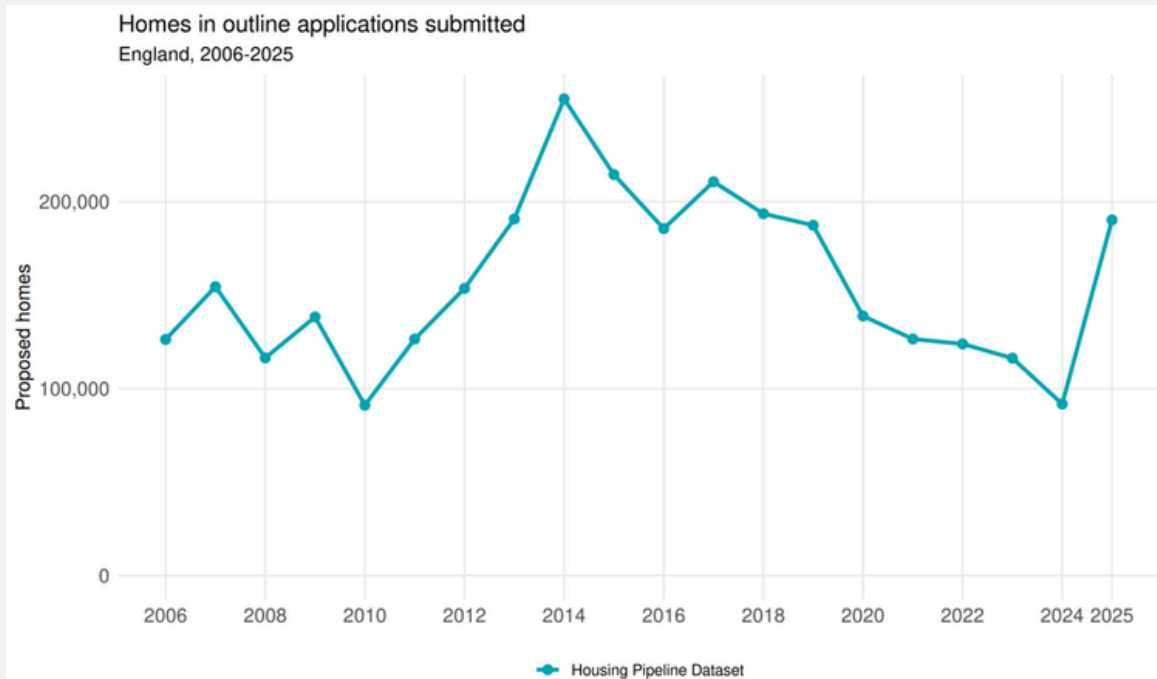
- Homes proposed through outline applications peaked in 2014 and declined over the following decade.
- Following the National Planning Policy Framework (NPPF) reform in December 2024, there was a 107% increase year-on-year in proposed homes. This increase does not by itself establish the effect of the reforms.
- Counting proposed homes is harder than it appears. Planning records aren't usable data without proper cleaning:
 - Correcting inconsistent field labelling across local authorities, such as the type of application, affects around 1.5 million residential units.
 - Of 4.7 million unique application descriptions that appear at first sight to propose new housing, only 2.3 million actually relate to proposals for unique new housing supply.
 - Failing to group applications for the same development overcounts homes by around 46% relative to the project-level total.
- The Ministry of Housing, Communities and Local Government's (MHCLG) published figures are, on average, 42.6% higher than the Housing Pipeline estimates across the years compared. The series differ in their treatment of repeat applications and data processing.

Our Housing Pipeline dataset provides a detailed account of individual developments and their planning histories, allowing examination of both the scale of proposed housing supply and its progression through the planning system. It covers more than twenty years and 400 local authorities' planning histories.

It goes beyond standard application records contained on the application's planning portal page to incorporate the underlying documents submitted throughout the planning process, providing a substantially richer evidence base comprising over 120 terabytes of data.

Its construction has involved three years of collection and development, combining state-of-the-art machine learning and large language models with thousands of hours of manual validation and specialised knowledge of the planning process.

What the data show



The first output from the Housing Pipeline dataset covers homes in submitted outline applications for the past 20 years. The series removes all cases of double-counting when a project files multiple outline applications covering essentially the same units. Outline applications are the first substantive planning stage for around half of all proposed residential units and are more common with larger developments (e.g. more than ten units).

The data series shows that proposed residential units peaked in 2014 and generally declined over the subsequent ten years. Notably, between 2024 and 2025, the number of homes proposed through outline applications more than doubled. While this is not causal evidence, it's promising to see this uptick following the December 2024 changes to the NPPF.

Why the numbers are so difficult to get right

Planning portals contain a wealth of information, but were built to administer individual cases, not as a source of statistics for analysis. Their records do not provide a ready-made, consistent count of new homes. Recording conventions vary across authorities, over time and even between staff entering records in the same authority in the same year. These inconsistencies, alongside errors in basic fields and repeated applications for a single project, create substantial obstacles. The scale of these problems helps explain why published estimates can go wrong.

The research identifies several systematic issues which can have significant effects on our understanding of the housing supply chain:

- ▶ **Application type identifies what is being requested**, such as outline or full planning permission, approval of reserved matters or discharge of conditions. The raw records contain 14,879 distinct entries for a handful of actual application types.
- ▶ Decisions record the outcome of an application and **should contain only a handful of options**, such as “Approved”, “Rejected” and “Withdrawn”. However, authorities say these things in over 32,000 unique ways.
- ▶ **The same applies to many other important characteristics**. For example, application status records where an application stands in the administrative process, such as awaiting validation, under consideration or decided. Authorities record 76,554 distinct entries for this basic attribute.
- ▶ In the raw planning portal data, **application-type information was missing for around 832,000 applications** and required correction for a further 1.6 million.
- ▶ **Identifying new housing supply from a single-line application description is extremely difficult** in many cases. Housing references may concern alterations to existing homes, mixed-use schemes may obscure residential provision, and student accommodation or care homes may report bedspaces rather than dwellings, making simple counts highly sensitive to wording.
- ▶ **Several applications can concern the same homes**. A single development can appear repeatedly as proposals change and different phases proceed. In the database, around two in five project histories contain more than one full or outline permission. Simply adding the units recorded against individual applications produces an overcount of about 46%.

Case study:

A South Norfolk development, first applied for in 2013, appeared initially as a proposal for 385 homes, then as a revised scheme for 375, followed by phase applications for 89 and 246 homes and further amendments. These records span more than a decade but describe the evolving history of a single development. Unless these applications are recognised as belonging to the same project, the same homes can be counted repeatedly.

Data construction

The Housing Pipeline dataset draws on approximately 19 million planning applications across 400 authorities, spanning more than twenty years. It goes beyond standard application records to incorporate the underlying documents submitted throughout the planning process. Together, these comprise over 120 terabytes of data, providing a substantially richer account of proposed developments and the decisions affecting them.

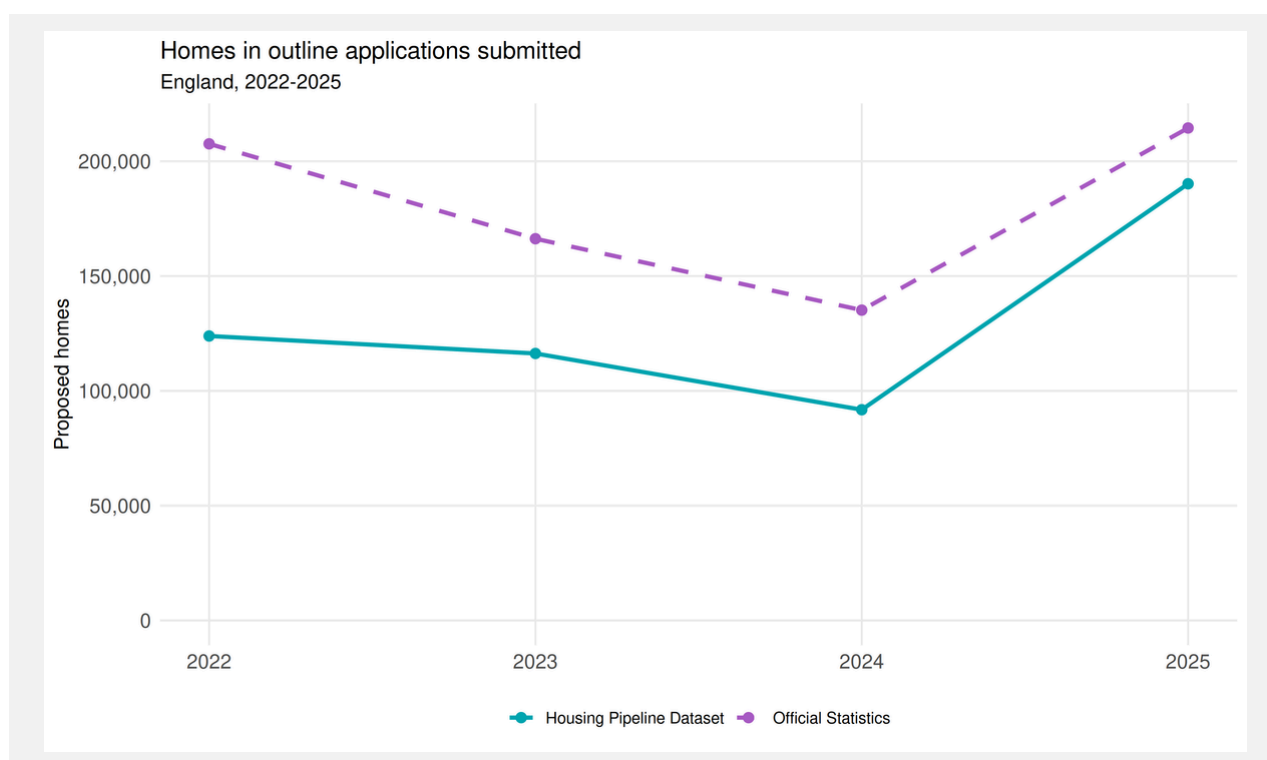
The three-year development process combined technical advances with detailed knowledge of planning. The dataset brings together bespoke machine-learning models, large language models, spatial and network analysis, and thousands of hours of manual human validation. These methods address different challenges including distinguishing proposals for new housing, identifying applications belonging to the same development, and recovering the number and type of homes proposed.

Accuracy is assessed against independently held-out examples. Our model for identifying applications proposing new housing achieves 99.3% accuracy on a held-out test sample. Alongside this, extensive manual classification and review underpin the interpretation of administrative fields. During development, we found that combining frontier AI tools with human validation and judgement produced the highest-quality dataset at the lowest cost. Neither approach alone offered sufficient quality at a reasonable cost.

The result is a detailed history of individual proposed developments. Approximately 3 million residential application records form 1.4 million development histories, covering around 10 million proposed residential units over 2000–2025. These histories retain the successive applications, decisions, revisions and phases associated with a project, allowing changes in its scale and progression to be examined over time.

How do the numbers compare with figures published by government?

Figures published by the MHCLG are, on average, 42.6% higher than the Housing Pipeline estimates each year over 2022–2025. Interestingly, the Housing Pipeline dataset finds a more sizeable uptick in proposed homes from outline applications following the 2024 NPPF reform. The Housing Pipeline dataset finds a 107% increase, while the published statistics from MHCLG find only a 59% uptick.



Several factors may explain the differences between the two measures.

Different treatment of repeat applications

MHCLG removes duplicates within a 12-month window, whereas Housing Pipeline links repeat applications across years within projects. Revised or replacement applications for the same homes may therefore contribute to MHCLG's figures in more than one year, whereas Housing Pipeline iteratively revises numbers when replacement applications occur. As such, the MHCLG series is an indication of wider planning activity in a given year and not an accurate reflection of the unique number of units applied for. Guidance from MHCLG suggests that its annual figures should not be aggregated across multiple years to estimate unique proposed homes whereas the annual Housing Pipeline figures can be aggregated in this way.

Data Processing

Double counting may occur if applications relating to the same project are not correctly linked. As described above, inconsistencies in planning records can make this difficult, especially without using details contained in the underlying documents. Additionally, issues around classification and interpretation of records as discussed in the previous section may also create discrepancies.

MHCLG Published Limitations

MHCLG describes the applications series as provisional management information providing a broad indication of trends. Its technical documentation acknowledges that some double counting remains despite filtering, and that figures are subject to revision as the source data (a data feed purchased from Glenigan) is updated.

Conclusion and future work

The new Housing Pipeline dataset can fundamentally improve our understanding of proposed housing supply, strengthen the available evidence and improve the data landscape. The full Housing Pipeline Outline dataset, released along with this brief, provides annual figures on outline applications and residential units going back twenty years. Future analysis will document the rates at which proposals progress to completed homes. Additionally, the full pipeline records will allow us to investigate where housing is proposed, how long applications take, how schemes change, where proposals encounter delays or refusals and subsequently, what gets built.

The underlying documents provide further evidence on planning requirements and the reasons behind decisions. This is the first in a series of briefs drawing on the dataset to examine the planning process and its role in housing supply.

Contact the researchers

Nikhil Datta, University of Warwick: nikhil.datta@warwick.ac.uk

Amrita Kulka, University of Warwick: amrita.kulka@warwick.ac.uk