# TESTING LAND VALUATION METHODOLOGIES

FINAL REPORT

**Lot 1: Market-based statistical valuation**
**Lot 2: Advanced algorithmic and machine-learning applications**
**Lot 3: Formula-based valuation by land area**
**Lot 5: Innovative or experimental approaches**

Dr Rhys ap Gwilym, Prof Adrian Gepp, Dr Sadeque Hamdan, Dr Edward Jones, Pretty Karibo, Dr Graeme Pearce, Temidayo Popoola, Dr Xiaoxi Qu

Bangor University

March 2026


PRIFYSGOL BANGOR UNIVERSITY

Ysgol Fusnes Albert Gubay
Albert Gubay Business School

## EXECUTIVE SUMMARY

The Welsh Government commissioned this project to test, compare and understand a range of methodologies for estimating land values in Wales. The work does not attempt to identify a single "true" value for land. Instead, it assembles the most comprehensive Wales-wide land, property and amenity datasets yet brought together, applies multiple modelling approaches at national and local levels, and examines how members of the public respond to different valuation methods. The findings show what can be achieved with today's data, what cannot, and where future effort would most usefully be directed.

### WHAT WE DID

We developed and assessed four complementary strands:

- **Lot 1 – Hedonic regression.** This approach uses large numbers of past property sales to estimate how different characteristics — such as location, plot size, property type, proximity to services and local amenities — contribute to observed sale prices. Hedonic models are well-established and widely used in economic analysis, and they provide a structured way to separate the value of land from the value of buildings**.**
- **Lot 2 – Machine-learning modelling.** This strand applies more flexible and modern modelling techniques capable of capturing complex, non-linear relationships that traditional models may miss. These methods generally deliver the highest predictive accuracy for overall property, but their internal workings can be less intuitive to interpret.
- **Lot 3 – Formula-based approaches.** This strand reviews how other countries develop and maintain land-value systems, and tests simplified formula-based approaches inspired by international practice. These systems emphasise predictability, stability and ease of communication. They typically assign values using area-based factors or zoning rules rather than detailed parcel-level modelling. They are best viewed as simplified frameworks, not stand-alone valuation tools.
- **Lot 5 – Behavioural evidence.** Lot 5 explores how members of the public interpret and respond to different valuation methods. Using an interactive dashboard and small financial incentives, participants compared parcel-level outputs, examined underlying data, and expressed preferences among modelling approaches. The findings suggest that participants perceive a trade-off between accuracy and interpretability of land valuation methods. This strand highlights the importance of public acceptability and transparency if a valuation system were ever to be used in practice.

All lots draw on a unified data backbone created for the project: a Transactions Database (1.28 million Wales transactions, scaled to 2025 values), a National Land Parcel Database (1.41 million polygons with spatial attributes), and an LSOA Land Parcel Database for nine diverse case-study LSOAs. Together they enable like-for-like comparisons across methods and geographies.

## KEY FINDINGS

**Data is the binding constraint; modelling is secondary**

- The project successfully fuses multiple data sources and constructs new spatial attributes. Nevertheless, critical gaps and inconsistencies remain – parcel geometries are problematic, coverage is incomplete, matching properties across datasets is challenging and there is a lack of data on key variables such as the planning system and the quality of amenities.
- These issues limit all modelling approaches. They explain why methods that should, in principle, be highly accurate still fall short of "policy-grade" precision for many uses. Improving the core data spine will yield larger gains than switching modelling techniques.

**Three modelling approaches; three distinct strengths**

- **Hedonic (Lot 1).** Provides an explicit, economically coherent decomposition of land and structure. Predictive performance is respectable for residential transactions, with coefficients that behave sensibly and a statistically insignificant intercept – an important signal that the assembled land-related variables are doing real work rather than relying on a baseline constant.
- **Machine learning (Lot 2).** Delivers the lowest prediction errors for total property values and captures important non-linearities and interactions that linear models cannot. It demonstrates what best-possible accuracy looks like given today's data, and produces spatial results that closely mirror Lot 1 at aggregate scales.
- **Formula-based (Lot 3).** Intentionally simple and communicable. Produces plausible relative spatial patterns using MSOA fixed effects, but requires external calibration to estimate absolute land values.

Across Lots 1–3, the broad geography of land values is consistent: the lowest values are concentrated in the south Wales valleys and other post-industrial areas; the highest values are found in parts of Cardiff, Monmouthshire, Swansea and selected coastal or amenity-rich pockets. This convergence suggests the underlying spatial structure of land value in Wales is robust to modelling choices, even if levels differ.

**Decomposition is feasible in all models, but never incontrovertible**
All three modelling approaches decompose land and structure by design. However, because “pure land value” is rarely and unreliably observed in the market, no decomposition can be uniquely validated. The philosophy matters: land and structures are complementary; separability is a modelling assumption, not an empirical truth. Lot 2’s strongest performers – allowing richer interactions – reinforce this point.

**Public preferences emphasise understandability**
Lot 5 demonstrates that participants consistently preferred the hedonic model, followed by machine learning and then formula-based approaches. Crucially, these preferences do not change under different financial incentives. This reinforces the importance of approaches that people feel they can understand and that appear grounded in recognisable logic, even if those approaches are not the most statistically accurate.

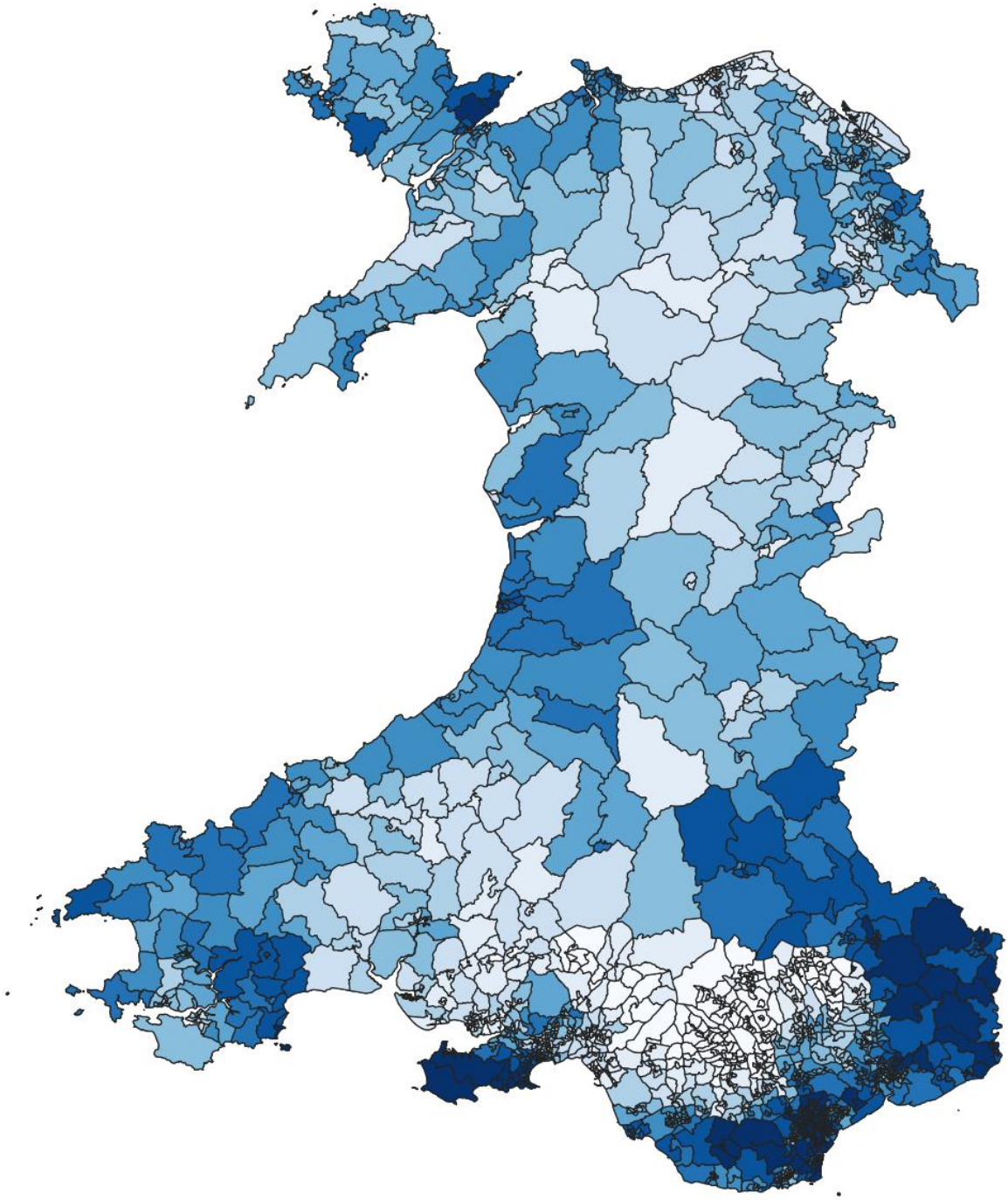
## WHAT THE MAP SHOWS (LOT 2)

Figure ES-1 presents average estimated land values per parcel in the National Land Parcel Database by LSOA using our best-performing Lot 2 configuration. It is included because machine-learning methods achieved the strongest predictive accuracy for property values and produce spatial variation that is very similar to the hedonic model. The map should be read as a plausible, data-constrained portrait of Wales’ land-value geography – not a definitive measure of the “true” value of land.

## CROSS-CUTTING LESSONS

1. **Land value is a construct, not an observed quantity.** Without systematic sales of unimproved land, all land values are model-based inferences. Different policy purposes imply different conceptual definitions (residual for insurance; social/location value for fiscal instruments; development potential for planning).
2. **Modelling discretion is unavoidable – its location differs.** Hedonic models require explicit choices about functional form and variables; machine-learning models automate many such choices inside algorithms; formula-based approaches place discretion with the formula’s architects and calibration choices. No approach “removes” judgement.
3. **Data quality dominates method choice.** With better core data (parcel–property linkages, planning designations, amenity quality, non-residential structures), all approaches improve—especially their usability for policy.
4. **Spatial patterns are robust, levels are not.** Methods agree on where values are high or low; they differ more on how high or how low.
5. **Public legitimacy depends on intelligibility.** Understandable methods, transparent assumptions and visible data quality improve acceptability.

**Figure ES-1: Lot 2 average land values per parcel in the National Land Parcel Database by LSOA**



Darker shades of blue represent higher land values per parcel.

*(Reproduces Figure 3.2.2 in the main report.)*

Links to interactive, scrollable maps: Lot 1, Lot 2, Lot 3.

Links to maps at the parcel-level for the nine identified LSOAs: Lot 1, Lot 2, Lot 3.

## WHAT THIS MEANS FOR POLICY AND IMPLEMENTATION

- **Feasibility today.** All three modelling approaches are operationally feasible (we implemented each), but none delivers accuracy that would be acceptable for high-stakes statutory uses across Wales with current data.
- **Role of models.** Methods can be combined:
    - Hedonic models offer explicit control and a coherent decomposition.
    - Machine-learning models are well-suited to benchmarking, diagnostics and revealing interactions; they may be central if the chosen definition of land value requires non-linear complementarities to be modelled explicitly.
    - Formula-based outputs can support communication and stability when calibrated by richer models.
- **Institutional data spine.** The primary constraint is not modelling but data governance. Fragmented responsibilities and inconsistent identifiers make reliable parcel-to-property-to-attribute linkage difficult. Any durable valuation system will require a more coherent Welsh land information infrastructure.
- **Public engagement as a data asset.** The Lot 5 dashboard shows people can and will help validate and correct local parcel information. A controlled feedback mechanism could provide a distributed quality-assurance layer for Welsh land and property data.

## SUMMARY

This project shows that land valuation in Wales is both possible and inherently constrained. The modelling work demonstrates clear progress on previous studies and strong internal coherence, but the accuracy of estimated land values is limited by data that is incomplete, inconsistent and not designed for the purpose.

At the same time, the project reveals stable spatial patterns across methods and provides new insight into how people perceive the legitimacy of valuation approaches.

Further progress will depend on clarifying the policy purpose of “land value”, improving the core data infrastructure, and governing modelling choices transparently – ideally with citizens helping to validate the data that describe their own places.

# CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## GLOSSARY

| Term/Acronym | Definition |
|---|---|
| API (Application Programming Interface) | A set of rules and tools that allows different software applications to communicate and exchange data. |
| Bayesian Information Criterion (BIC) | Model selection criterion that penalises complexity to favour parsimonious models. |
| BLPU (Basic Land and Property Unit) | A type of UPRN in the OS AddressBase database. |
| British National Grid | UK coordinate reference system used for mapping and distance calculations. |
| CCOD | Commercial and Corporate Ownership Data – HM Land Registry dataset of commercial property ownership. |
| Cross-validation | Technique to estimate out-of-sample performance by training/testing on multiple folds. |
| CSV (Comma-Separated Values) | A simple, plain text file format for storing tabular data (like spreadsheets or databases) where each line is a data record and values within a record are separated by commas. |
| DAFI | Data Analytics and Financial Innovation – Bangor University research group. |
| DPA (Delivery Point Address) | A type of UPRN in the OS AddressBase database. |
| EPC (Energy Performance Certificate) | Source of building-level data on energy efficiency and structural characteristics. |
| FEMA Retail Areas | Ordnance Survey Functional Economic Market Area classification of retail centres. |
| GAM (Generalised Additive Model) | Flexible regression that allows non-linear effects via smooth functions. |
| Geospatial analysis | Analysis of data with explicit geographic/locational components using GIS methods. |
| GIS (Geographic Information System) | Computational tools that capture, store, manage, analyse, and display all types of geographically referenced data. |
| GP | General Practitioner – primary care medical doctor. |
| Hedonic regression | A method that models property prices as a function of structural and locational attributes to infer land value. |
| HM Land Registry | A non-ministerial department of the UK Government, responsible for registering the ownership of land and property in Wales and England. |
| INSPIRE Index Polygons | HM Land Registry polygon boundaries for registered land parcels. |
| KD-tree | A space-partitioning data structure that organizes points in a k-dimensional space, acting like a binary search tree for multiple dimensions. |
| kNN (k nearest-neighbour) | Method that finds the closest feature(s) in space to compute distances or impute values. |

| | |
|---|---|
| Lasso (Least Absolute Shrinkage and Selection Operator) regression | A statistical method that performs both regularization and variable selection by adding a penalty term to the cost function, which shrinks the regression coefficients. |
| LDP (Local Development Plan) | The key document each local authority creates to guide future land use and development (housing, jobs, green spaces) for about 15 years. |
| LSOA | Lower Layer Super Output Area – small geographic unit used for statistics in the UK. |
| LVT | Land Value Tax – a tax based on the value of land rather than property structures. |
| MAE (Mean Absolute Error) | Average absolute difference between predictions and actual values. |
| MSE (Mean Squared Error) | Average squared difference between predictions and actual values. |
| NaPTAN | National Public Transport Access Nodes – dataset of public transport stops and stations. |
| NN (Neural Network) | Machine learning model that captures complex, non-linear relationships. |
| NRW (Natural Resources Wales) | A Welsh Government Sponsored Body, whose purpose is to ensure that the natural resources of Wales are sustainably maintained. |
| NSW | New South Wales – Australian state referenced in international examples. |
| ONS (Office for National Statistics) | UK's statistics authority. |
| OS (Ordnance Survey) | The mapping agency for Great Britain. |
| OS MasterMap | Ordnance Survey high-resolution topographic dataset. |
| OSM (OpenStreetMap) | Crowd-sourced geospatial database of roads, buildings and points of interest. |
| PDF | Portable Document Format – document file format. |
| PPD | Price Paid Dataset – HM Land Registry dataset of property transactions. |
| RF (Random Forest) | Ensemble of decision trees for regression/classification. |
| RMSE (Root Mean Squared Error) | Square root of average squared prediction error – measure of prediction accuracy. |
| Stepwise regression | Procedure that adds/removes variables to find a well-fitting, parsimonious model. |
| SVR (Support Vector Regression) | Margin-based regression algorithm. |
| UPRN (Unique Property Reference Number) | Unique identifier for every addressable location in the UK. |
| WIMD (Welsh Index of Multiple Deprivation) | Measure of relative deprivation for small areas in Wales. |
| WRA (Welsh Revenue Authority) | A non-ministerial department of the Welsh Government responsible for the administration and collection of devolved taxes in Wales. |

# 1. INTRODUCTION

The Welsh Government commissioned this project to examine how different methods for estimating land value perform when applied in a Welsh context. Using newly integrated datasets and a programme of modelling across nine contrasting areas, the report assesses the strengths and limitations of each approach for potential future policy use.

## THE RESEARCH TEAM – EXPERTISE AND CREDENTIALS

This project brought together Bangor University's Tax and Welfare and Data Analytics and Financial Innovation (DAFI) research groups. The Tax and Welfare Group has a long track record of working on tax research projects in a Welsh context. Notably, this includes the only previous detailed study of land values in Wales (ap Gwilym et al 2020). The DAFI Group has expertise in applying the latest methods in data analytics to various policy and business contexts.

Dr Rhys ap Gwilym – Project Manager

- PhD Economics (Cardiff University).
- Senior Lecturer in Economics, expert in Regional Economics, Public Finance, and the Welsh Economy.
- Recognised expertise on land valuation (ap Gwilym et al 2020).

Professor Adrian Gepp – Quality assurer

- PhD in Applied Statistics (Bond University).
- Expert in predictive modelling, fraud detection, and valuation methodologies.
- Holds £600,000+ in external research funding and serves on editorial boards of international journals.
- Fellow of the Royal Statistical Society.

Dr Sadeque Hamdan – Lead on machine learning

- PhD Complex Systems Engineering (University of Paris Saclay).
- Senior Lecturer in Data Analytics at Bangor University.
- Published in top journals in transportation management and operational research.
- Award-winning researcher with expertise in Geographical Information Systems and the collection, processing and modelling of spatial/mapping data.

Dr Edward Jones – Lead on hedonic modelling

- PhD Economics (Bangor University).

- Senior Lecturer in Economics, specialising in quantitative modelling and financial economics.
- Fellow of the Royal Statistical Society and member of the Chartered Management Institute.
- Expertise in hedonic modelling of land values (ap Gwilym et al 2020).

Pretty Karibo – Lead on GIS analysis, spatial data engineering, and digital platform development

- PhD candidate in Social Policy at Bangor University, applying advanced data analytics and machine-learning methods to policy research.
- IBM-certified Data Scientist with expertise in Python development, applied modelling, and automated data processing.
- Specialist in geospatial analysis and the visualisation of complex land and amenity datasets.

Dr Graeme Pearce – Lead on experimental design

- PhD Economics (University of Exeter).
- Senior Lecturer in Economics, specialising in behavioural and experimental economics.
- Published widely in top economics, management and inter-disciplinary journals.
- Extensive expertise in behavioural economics and experimental methodology and design.

Temidayo Popoola – Lead on data integration

- MSc Data Analytics.

Dr Xiaoxi Qu – Literature review and econometric modelling

- PhD Finance (University of Chinese Academy of Social Sciences).
- Visiting researcher at Bangor University.

## BACKGROUND AND CONTEXT OF VALUING LAND IN WALES

In the Welsh context, two of the members of our team (Rhys ap Gwilym and Edward Jones) were among the authors of the only previous detailed study of land values in Wales (ap Gwilym et al 2020). Our work considered the potential for implementing a local LVT in Wales to replace council tax and/or non-domestic rates. As part of this assessment, we carried out detailed modelling of land values in Wales, covering all land underlying properties that were liable for council tax or non-domestic rates in 2019. We published average land values at the LSOA level based on estimates produced by hedonic regressions at the individual property level.

We concluded in that report that the main challenges in undertaking land valuation in Wales relate to data availability, especially property characteristics for non-residential land, and transactions data for unimproved land. The modelling techniques themselves are well-established and relatively straight-forward to implement.

## INTERNATIONAL EXAMPLES OF VALUING LAND

Approaches to valuing land vary widely across jurisdictions, reflecting differences in legal frameworks, data availability, and policy objectives. Broadly, three types of systems are observed internationally:

**Area-Based Models:**

In several countries, land value or tax liability is calculated using only parcel size and fixed statutory amounts. Examples include Bavaria's Flächenmodell in Germany and similar systems in Poland, Romania, and Slovakia, where the formula is essentially land area × fixed amount per m², sometimes adjusted by simple coefficients.

**Value-Based Systems Using Mass Appraisal:**

Countries such as Denmark, Estonia, Australia, the Netherlands, and South Africa rely on assessed land or property values derived from market-based models. These systems incorporate parcel area, location, and land-use category into hedonic or comparable-sales frameworks to produce statutory values.

**Hybrid or Specialised Models:**

Some systems combine area-based simplicity with market or productivity adjustments. Germany's federal Bundesmodell uses parcel area × official standard land value (Bodenrichtwert), while Japan's Rosenka system applies roadside unit values with correction factors.

These examples illustrate the spectrum from highly transparent, formula-driven models to complex valuation systems grounded in market evidence. A detailed review of these approaches and their implications for Wales is provided in the Lot 3 section of this report.

## THE NINE IDENTIFIED LSOAS

The project brief specified nine Lower Layer Super Output Areas (LSOAs) to enable detailed parcel-level modelling and to illustrate how the valuation approaches behave in contrasting geographic contexts. Table 1 provides a brief overview of these LSOAs, and Figure 1 shows their location.

**Table 1: Overview of the Nine Identified LSOAs**

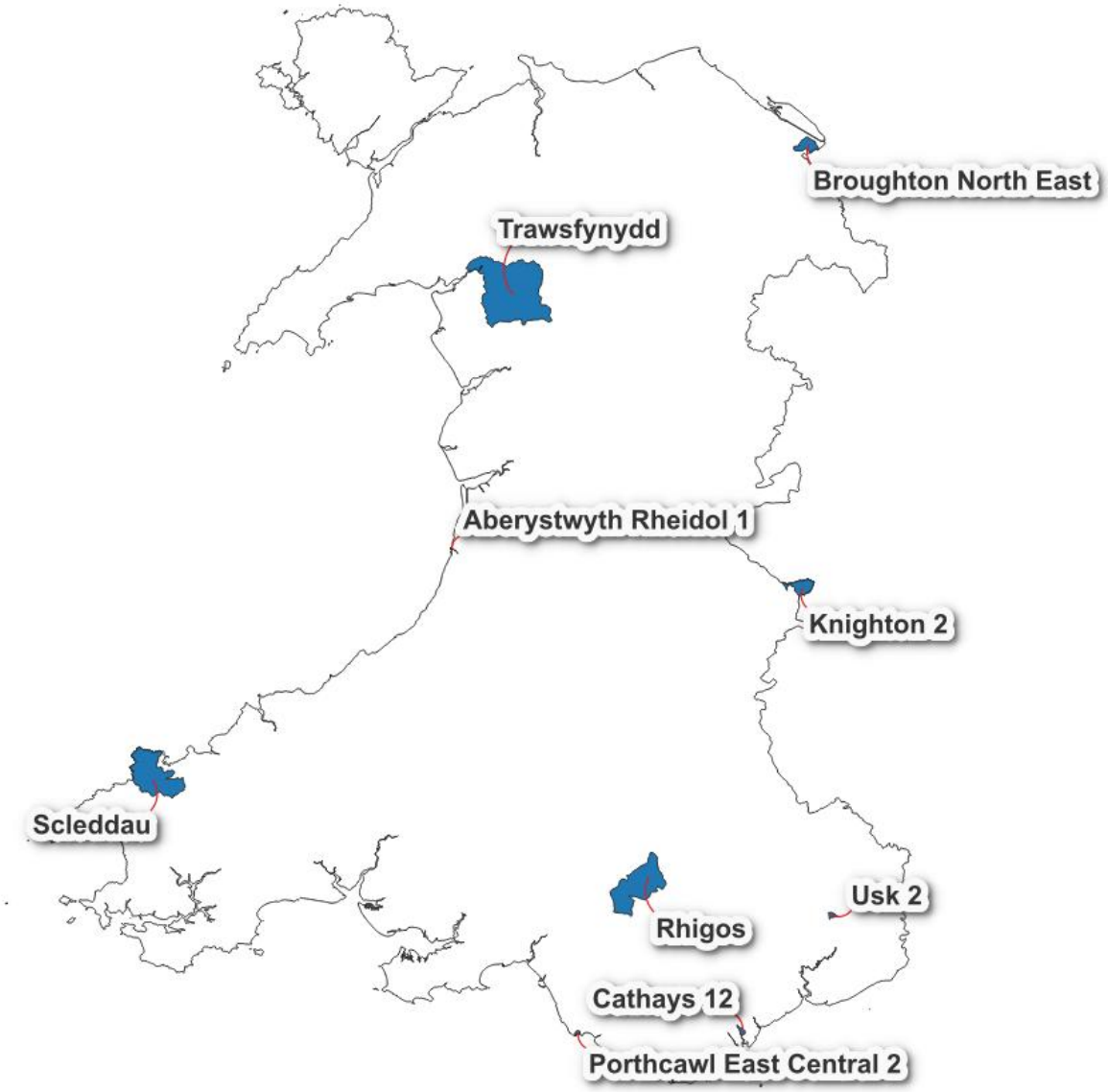
| LSOA code & name | Brief description | Population | Area / $km^2$ | WIMD 2025 rank |
|---|---|---|---|---|
| **W01000114 – Gwynedd 009D (Trawsfynydd)** | Rural village and surrounding upland area | 1,397 | **163.5481** | **796** |
| **W01000255 – Flintshire 015A (Broughton North East)** | Suburban area adjacent to industrial and retail zones | 2,320 | **9.2111** | **1464** |
| **W01000449 – Powys 011C (Knighton 2)** | Small market town and rural fringe | 1,525 | **11.8339** | **1376** |
| **W01000517 – Ceredigion 002D (Aberystwyth Rheidol)** | Mixed university, residential and commercial area in central Aberystwyth | 1,154 | **0.2751** | **744** |
| **W01000617 – Pembrokeshire 002F (Scleddau)** | Rural, coastal hinterland west of Fishguard | 1,543 | **71.4865** | **636** |
| **W01001045 – Bridgend 019D (Porthcawl East Central 2)** | Coastal town centre area including mixed residential & tourism zones | 1,697 | **1.1537** | **684** |
| **W01001233 – Rhondda Cynon Taf 001F (Rhigos)** | Former coalfield villages; dispersed communities in upland valley | 1,803 | **74.1898** | **764** |
| **W01001597 – Monmouthshire 006F (Usk 2)** | Historic small town and residential area | 1,365 | **1.4811** | **1824** |
| **W01002019 – Cardiff 032H (Cathays 12)** | Cardiff city centre. Retail areas, parks and civic centre | 3,366 | **1.7315** | **1062** |

Sources: Population – ONS mid-year 2023 population estimates

Area – ONS Standard Area Measurements for 2021 Statistical Geographies

WIMD 2025 – StatsWales

**Figure 1: Location of the nine identified LSOAs**



# 2. METHODOLOGY

This project has been delivered across four interconnected lots, each examining a different approach to land valuation in Wales: a hedonic pricing model (Lot 1), machine-learning valuation (Lot 2), formula-based methods (Lot 3), and an experimental investigation of citizen preferences over land valuation methods (Lot 5). Although each lot has a distinct methodological focus, all four share a common data foundation. In particular, we constructed three unified, all-Wales datasets – the Transactions Database, the National Land Parcel Database, and the LSOA Land Parcel Database – which provide a consistent and robust empirical basis for every element of the analysis. The project also includes more detailed work on nine focus LSOAs specified by the Welsh Government. These areas represent a diverse set of economic, geographical and settlement contexts across Wales, and are used throughout the report to demonstrate the performance, limitations and practical implications of each valuation method at a fine spatial scale. Together, this structure ensures that the work is both methodologically rigorous at the national level and grounded in detailed, place-specific evidence where it matters most for policy.

## LOT 1: MARKET BASED STATISTICAL VALUATION

Lot 1 uses a hedonic regression framework to decompose observed property transaction prices into their underlying land and structure components. This approach follows the regression-based residual method described in the academic land-valuation literature (see Zhou et al., 2025) and extends the methodology employed in ap Gwilym et al. (2020).

The economic premise of the approach is that a property's value reflects the combined contributions of (i) land-related attributes, such as location, accessibility, environmental characteristics and neighbourhood qualities, and (ii) structure-related attributes, such as building area, type, age and configuration. By estimating these contributions separately, it becomes possible to recover an implied land value even in the absence of reliable market transactions for unimproved land – a particularly important consideration in Wales, where such transactions are extremely rare and often idiosyncratic.

Hedonic regression extends the basic regression framework by modelling property prices as the combined value of their structural features and location-related factors. First formalised by Rosen (1974), the approach assumes that the price of a property reflects the implicit value that buyers assign to both the structure itself and the attributes of the land on which it sits. By estimating how each characteristic contributes to the overall price, hedonic models allow us to decompose observed transaction values into the separate effects of structural features and location-related

factors. This makes hedonic regression particularly suitable for estimating land values in Wales, where transactions of unimproved land are limited, and the contribution of land must therefore be inferred indirectly from the behaviour of the property market.

In housing and land markets, buyers do not purchase land for its own sake but for the characteristics associated with a specific location (Malpezzi, 2003). These include access to services, environmental quality and neighbourhood features, all of which shape how households and businesses value different places. Because land is fixed in space, choosing a property means choosing a particular bundle of these characteristics. Hedonic pricing builds on this idea, treating market prices as the outcome of buyers competing for locations that best match their preferences and constraints (Quigley, 1985). Modern empirical research reinforces this view, showing how structural features, neighbourhood amenities and spatial accessibility are systematically capitalised into property values (Ahlfeldt and Wendland, 2009, and Gibbons et al, 2014). The hedonic equation therefore describes how property prices vary with these underlying structural and locational attributes, providing a rigorous framework for estimating the value of land in different contexts.

## MODEL SPECIFICATION

The dependent variable in the Lot 1 model is the scaled price paid.[1] $H_i$ is the original transaction price uplifted to 2025 values using official house-price indices (see data section for details). This inflation adjustment allows the model to be estimated without time dummies (which performed poorly due to the large number of required dummy parameters relative to the number of observations).

The core hedonic specification is:

$$H_i = \alpha + \beta^L \cdot x_i^L + \beta^S \cdot x_i^S + \epsilon_i \qquad (1)$$

where:

- $x_i^L$ denotes land-related attributes,
- $x_i^S$ denotes structure-related attributes,
- $\alpha$ is a constant term, whilst $\beta^L$ and $\beta^S$ are vectors of estimated coefficients,
- and $\epsilon_i$ is an idiosyncratic error term.

Land-related attributes include parcel area, distances to amenities, transport accessibility, local environmental factors, topography, neighbourhood socio-economic indicators and other spatial variables. Structure-related attributes

[1] In practice, we estimated several specifications of Equation 1, including versions where the dependent variable was expressed in levels and in natural logs, and where covariates were normalised or transformed. The choice of functional form affects performance but not the underlying conceptual decomposition.

include building characteristics, building-footprints, property type and age, and other attributes. These are explained in detail in the data section below.

The model is estimated on Wales-wide transaction data, consistent with best practice for hedonic land-value estimation (e.g. Kuminoff et al., 2010), because the nine identified LSOAs lack the transaction volume and attribute heterogeneity required to estimate stable coefficients.

## ESTIMATING LAND VALUE

Following the regression-based residual method (Zhou et al., 2025), the implied land value for property $i$ is defined as the component of the predicted log price attributable solely to land-related attributes:

$$\widehat{\ln(L_i)} = \theta\hat{\alpha} + \hat{\beta}^L \cdot x_i^L$$

Where $\theta$ is the proportion of the constant term that represents land values. In a well-specified model, the observable land and structure-related attributes capture all the systematic variation in log prices without requiring a substantial baseline shift. Hence, the constant term $\hat{\alpha}$ will not be statistically different from zero, and $\theta\hat{\alpha}$ is therefore equal to zero. In this case, we can take the natural exponent of the previous term to obtain an estimate of the land value in levels:

$$\widehat{L_i} = \exp\left(\hat{\beta}^L \cdot x_i^L\right) \qquad (2)$$

Because the intercept $\hat{\alpha}$ is statistically insignificant, there is no practical ambiguity about how to apportion it across land and structure components. This simplifies the decomposition substantially: all systematic land value is captured through $\hat{\beta}^L \cdot x_i^L$, and all systematic structure value through $\beta^S \cdot x_i^S$.

This method is particularly appropriate in Wales, where the number of transactions involving unimproved land is extremely small, and where those few transactions that do occur often involve atypical or tax-advantaged parcels that are not representative of market-wide land values. The regression-based method avoids relying on these sparse and noisy data sources.

## MODEL EVALUATION AND DIAGNOSTICS

The performance of the hedonic regression model is assessed using standard evaluation metrics that indicate how well the model explains variation in property prices and how accurately it predicts observed values. The primary measure of explanatory power is the coefficient of determination ($R^2$), which shows the proportion of variation in log price per square metre that is accounted for by the included structural, locational and environmental characteristics. A higher $R^2$ indicates that the model captures more of the relevant market dynamics, although it

is interpreted alongside other diagnostics. Predictive accuracy is evaluated using the Root Mean Squared Error (RMSE), which summarises the average difference between the model's predicted prices and actual market prices. RMSE is expressed in the same units as the dependent variable and is particularly useful for comparing the performance of alternative model specifications. Together, $R^2$ and RMSE provide a balanced assessment of how well the model fits the data and how reliably it can be used for estimating land values across Wales.

Since direct market evidence on unimproved land is essentially absent, external validation relies on comparison with the estimates in ap Gwilym et al. (2020), internal consistency checks against the Lot 2 machine-learning model, spatial plausibility checks at the parcel and LSOA level.

## SUMMARY

The Lot 1 hedonic methodology provides a transparent, statistically rigorous and economically grounded means of estimating land values across Wales. The approach delivers parcel-level land-value estimates without relying on rare and unrepresentative transactions of unimproved land. This provides the foundation against which the alternative methodologies explored in the other lots can be compared.

## LOT 2: ADVANCED ALGORITHMIC AND MACHINE-LEARNING APPLICATIONS

Our approach prioritises interpretability over black-box solutions. Additionally, we enable comparisons between statistical and machine learning approaches that are lacking in the literature.

Comparative studies have found (extreme) Gradient Boosting to outperform a neural network and random forest (Ma et al 2020; Jafary et al 2024), and Zhang et al. (2021) found encouraging results with extra trees regression and a radial basis function-based Support Vector Regression (SVR). However, training such models requires a large dataset of reliable historical land values that is unavailable here. Whilst the annual number of property transactions in Wales is in the tens of thousands (50,160 in 2023/24 according to StatsWales, 2025), the number of transactions of unimproved land is in the low hundreds.

We instead need machine learning to decompose property values into their building and land components, as our Lot 1 hedonic approach does by, in simple terms, estimating

$$H_i = \beta^L \cdot x_i^L + \beta^S \cdot x_i^S + \epsilon_i \qquad (3)$$

where $H_i$ are property values, $x_i^L$ are land-related attributes (location, distance from amenities etc), $x_i^S$ are structure-related attributes (number of rooms, property type, year of construction etc) and $\epsilon_i$ is an error term.[2] The land component is estimated as $\beta^L \cdot x_i^L$ because it is independent of structure ($\beta^S \cdot x_i^S$). This independence is not the case with the machine learning methods above, as they model interactions between the $x$ variables. While it is possible to train such models for property values and then set the structure-related variables to 0 to get land estimates, this would be an unreliable extrapolation that is methodologically unsound because of the interactive effects.

Our **machine learning enhanced hedonic approach** has the potential to improve the modelling whilst retaining the final model transparency.

- Generalising equation 3 to handle varying property types (residential, non-domestic, agricultural), necessitates interactions between property type and land variables. Decision trees, and some Bayesian networks, will guide the decision of which interaction terms to include – this approach is validated by our research in financial literacy (Xue et al. 2019). Furthermore, the tree will reveal any need for

[2] Equation 3 expresses the conceptual requirement that machine-learning methods should, like the hedonic model in Equation 1, produce separate land- and structure-related components. In practice, the ML models do not enforce this separability: they learn complex interactions among variables, and their internal representations include non-linearities and algorithm-specific “bias” terms rather than a conventional regression constant. For the purpose of comparison across Lots, we therefore express the ML decomposition in the same additive form as Equation 1, while recognising that this is a conceptual illustration rather than a literal description of the underlying model.

interactions between structure or between land variables (but not between both to retain the ability to decompose structure and land).
    - This approach will also be compared with separate modelling for the three property types separately.
- A lasso regression framework will be used for Equation 3 because it performs both variable selection and regularisation, which are important given we have such a large dataset that risks us finding spurious relationships if all variables are included.
- Equation 3 assumes a fixed linear relationship $\beta \cdot x$, but it is possible that some variables have a non-linear effect better suited to a more flexible structure $f(x)$ using a spline. Using a Generalised Additive Modelling (GAM) approach that retains the ability to decompose structure/land, non-linear effects will only be introduced if there is evidence in the data to ensure models are not unnecessarily complex (a risk with modern analytics).

Wales-wide data will be used to develop models, consistent with Lot 1. The nine specified LSOAs had 6,901 households at the 2021 census. In our previous work, we found approximately 10% of residential properties had transaction values in Land Registry datasets. Therefore, the number of property transactions within these LSOAs alone is unlikely to enable the complex modelling we propose.

## BENCHMARK/ROBUSTNESS MODELS

1. Unlike radial basis function-based SVR, **linear SVR** is suitable for the separation into land and structure components and so will be used, consistent with prior research.
2. Complex models that fit well overall can still contain instances where two similar pieces of land have different valuations. Any difference can be explained using our models, but it might be a complex story not easily explained. **k-Nearest Neighbours (kNN)** avoids such instances, but is too simple to produce an accurate final model. Thus, kNN will be used to identify outliers in other estimates that will be reviewed, and then the underlying models adjusted as needed.

## EVALUATION

Root Mean Squared Error (RMSE) is commonly used in the machine learning literature. This metric is also shared with Lot 1, so its use enables comparisons with Lot 1 and those reported in ap Gwilym et al (2020). For Lot 2, 10-fold cross-validation will be used to create realistic error estimates as metrics from using the

entire training data are overly optimistic and not as representative of predicted future performance – 10-fold Cross validation is a common technique to address this.[3]

Additional evaluation of our approaches is made by fitting a powerful, flexible neural network (NN) and extreme Gradient Boosting (XGBoost) models to the data. This model will not allow us to separate out the land component as desired, but it will give us a better idea about what error levels are good. NNs are flexible function-approximators: they can capture highly non-linear relationships. XGBoost (a boosted tree method) is particularly effective for structured/tabular property data because it automatically learns strong "if-then" style rules and interactions.

We explore whether an XGBoost model fitted to total transaction prices can be used to recover an implicit "land-only" component via feature-attribution. After training XGBoost on the full dataset, we decompose each prediction into an additive baseline plus per-feature contributions, $\widehat{H}_i = b + \sum_j \phi_{ij}$, where $b$ is baseline (bias) and $\phi_{ij}$ is the feature $j$ contribution. We then partition features into land-related attributes (e.g., location, accessibility, neighbourhood metrics) and structure-related attributes (e.g., floor area, property type, construction characteristics), and form a pseudo land value by aggregating the contributions associated with land features, $\hat{L}_i = b + \sum_{j \in \mathcal{L}} \phi_{ij}$.

Because these pseudo land values inherit the model's non-linearities, we subsequently fit a second XGBoost model using only land-related variables to predict $\hat{L}_i$, enabling application to datasets where structure variables are absent (or intentionally excluded) while retaining an auditable mapping from land attributes to the inferred land component.

[3] A criticism of standard RMSE is that it is a measure of accuracy based on the data used to build the model, but if the model were to be implemented then we are concerned with how accurate the model will be on future data that the model has not seen before. 10-fold cross-validation is a way to address this problem and provide a better estimate of future accuracy. This process randomly partitions the data into 10 "folds", it then builds the model on 9 folds and then evaluates the resulting model's accuracy on the final 10th fold that is akin to future unseen data. This process is repeated 10 times so each fold is use for evaluation/test once; the accuracy on those 10 unseen folds is then aggregated to produce the 10-fold cross-Validated RMSE. It is akin to giving someone 9 quizzes to study, then giving them a 10th quiz to see how well they truly learned the content.

## LOT 3: FORMULA BASED VALUATION BY LAND AREA

Lot 3 examines the potential for developing simple, formula-based approaches to land valuation that could offer a more transparent and administratively lightweight alternative to the full statistical and machine-learning models developed in Lots 1 and 2. The original project specification envisaged two complementary strands:

- **Strand A:** a systematic review of international formula-based valuation systems, to identify the types of variables and structures used elsewhere; and
- **Strand B:** the development of simplified, parsimonious formulae for Wales by applying general-to-specific techniques to the richer models developed in Lots 1 and 2.

Both strands were completed. However, the work undertaken in Lots 1 and 2 revealed that the original approach for Strand B was not feasible in practice, and the findings from Strand A suggested that a direct modelling of international formula structures was also not meaningful in a Welsh context. As a result, the methodological approach for Lot 3 was adapted to reflect what is practically achievable with Welsh data and what is conceptually appropriate given international practice.

We followed two different approaches to identify parsimonious formulae for land valuation. We assessed the relative strengths and weaknesses of these formulae in terms of (1) simplicity and ease of understanding, and (2) precision in valuing land.

### STRAND A – SYSTEMATIC REVIEW OF INTERNATIONAL PRACTICE

We undertook a structured review of thirteen international jurisdictions where land valuation is supported by statutory or formula-based rules. These systems vary widely in complexity, purpose, and inputs. For each jurisdiction, we reviewed primary legislation, administrative guidance, and supplementary academic and policy literature.

Our extraction process focused on the following elements:

- the explicit variables appearing in statutory or formula-based valuation systems;
- whether land area enters the valuation formula directly or indirectly;
- the role of standardised unit values (such as Japan's *Rosenka*, Germany's *Bodenrichtwerte*, or Bulgaria's base values plus coefficients);
- the extent to which local authorities or expert bodies determine adjustments; and
- administrative and institutional requirements for maintaining such systems.

International systems were grouped into three broad types:

1. **Pure area-based models**, where land area is multiplied by a fixed statutory amount with limited or no spatial differentiation;
2. **Standardised unit-value models**, where a per-square-metre land value is set administratively for zones, streets, or land-use categories; and
3. **Full mass-appraisal systems**, where statistical or market-based models generate parcel-level land or property values.

The review offered valuable context but did not yield formula structures that are directly transferrable to Wales. Area-based systems are intentionally simplistic and offer little basis for modelling, while mass-appraisal systems replicate what is already achieved in Lots 1 and 2. Standardised unit-value systems often derive their values through expert-committee processes rather than replicable formulas, meaning the underlying methodology is not observable in a way that could be meaningfully applied to Welsh datasets.

All extracted information was compiled into structured tables summarising variables, formula components, and adjustment factors for each jurisdiction. These tables are presented in the findings for Lot 3.

## RATIONALE FOR REVISING THE STRAND B APPROACH

The original intention for Strand B was to derive simplified land-valuation formulae by applying general-to-simple principles to the rich hedonic and machine-learning models used in Lots 1 and 2. However, this approach proved impracticable for two reasons:

1. Limited scope for further simplification:
   Lots 1 and 2 already apply extensive variable selection, penalisation (LASSO), and interaction pruning. Further simplification either removes essential structure/land distinctions or collapses the model into trivial relationships (e.g., area alone), offering no useful insight.
2. Conceptual gap between international systems and available Welsh data:
   International “standardised value” models rely on administratively produced unit-value schedules (e.g. BRW zones, Rosenka maps). Wales has no equivalent institutional infrastructure, and these systems do not provide formulae that can be replicated statistically.

Given these constraints, a revised and more feasible approach was required for Strand B—one that is consistent with Welsh data, our previous work, and international practice.

## REVISED STRAND B: STRUCTURE-ONLY HEDONIC REGRESSION WITH MSOA FIXED EFFECTS

To produce a simple, formula-based valuation approach grounded in Welsh data, we adopted a method similar to that used in ap Gwilym et al. (2020).

We estimated a hedonic regression using only structure-related variables, with no parcel-level land or locational attributes. All locational effects – amenities, accessibility, neighbourhood characteristics, environmental context – are instead captured through MSOA fixed effects.

In this model:

- Structure variables (e.g., floor area, EPC attributes, property type) explain the value of the building component.
- MSOA dummies absorb all location-driven variation in prices.
- The MSOA coefficients therefore act as a "standardised land value index": a single location-specific number summarising the land component of value in that MSOA.

MSOAs were used to balance granularity and statistical robustness. This is a point of difference to ap Gwilym et al. (2020), which used the more granular LSOA units. However, given the richer set of structure attributes that we have derived in this project, it proved to be too computationally demanding to combine that set of attributes with LSOA level granularity.

The model takes the form:

$$\ln(H_i) = \alpha + \beta^L \cdot MSOA_i + \beta^S \cdot x_i^S + \epsilon_i \qquad (4)$$

where $H_i$ is the transaction price; $MSOA_i$ is a vector of binary variables denoting the MSOA in which the property is located; $x_i^S$ is a vector of structure-related attributes; and there are no land-related variables appear on the right-hand side.

The MSOA effects represent implicit land values, averaged across all transactions within each MSOA.

This parallels international "standardised unit value" systems while remaining fully replicable using Welsh administrative data.

## LOT 5: INNOVATIVE OR EXPERIMENTAL APPROACHES

Lot 5 uses a behavioural-experimental approach to examine how Welsh residents respond to different land valuation methodologies when these methods have financial consequences for them. The purpose of the exercise is not to validate the statistical performance of alternative models – that is the role of Lots 1–3 – but rather to understand whether members of the public have *systematic preferences* over different valuation approaches when those preferences are informed and consequential.

To elicit these preferences in a controlled environment, we designed and implemented a three-stage online experiment. The experiment also provided transparency around the underlying data, allowing participants to form informed preferences that might depend on their confidence in the data describing their own land parcel.

This aligns with the Welsh Revenue Authority's (WRA) principles of:

- **Cydweithio**: means 'to work together' and carries a sense of working towards a common goal.
- **Cadarnhau**: suggests a solid, robust quality that can be relied on. This is about providing certainty, being accurate and reinforcing trust.
- **Cywiro**: literally means 'returning to the truth' and is about the way they work with taxpayers to resolve errors or concerns.

In the spirit of 'Cydweithio', we engaged with individuals to assess their preferences over three potential land valuation methodologies. Participants are given the opportunity to confirm (i.e. 'Cadarnhau') or correct (i.e. 'Cywiro') the cadastral data that goes into those models. Importantly, the process provided a means to inform individuals of the meaning of 'land value' and its drivers, so as to promote a better understanding of land value as a basis for policy design, and potentially greater acceptance of policy reforms based on land value.

The experimental tax compliance literature (e.g. Cummings et al 2009, Fonseca and Grimshaw 2017) were used to inform the details of the surveys and dashboards that were developed. Following the economics literature, incentivised economic experiments were used to explore (1) individuals' attitudes towards land valuation and (2) how different incentives faced by the participants influences which land valuation methodology they most prefer. This second point of exploration seeks to understand how the desire for a *low* or *high* land valuation might influence the individual's preference for each land valuation methodology. This is a nuanced but important point, as many individuals may (for a variety of reasons) want their land to be valued at high or low levels.

We are using an experimental methodology because this affords us the ability to then make *causal* claims about how the incentives faced by individuals influences their preference for a specific land valuation methodology.

## EXPERIMENTAL DESIGN

All participants were recruited via the Prolific platform and completed the study using a bespoke web application built in Python/Flask with integrated geospatial visualisation. Both the recruitment and experimental interface ensured randomisation, consistent data handling, and replicable assignment to incentive treatments.

A total of **710 individuals** started the study. Because the experiment was limited to Welsh residents, the majority of exclusions occurred immediately at the postcode stage, with users entering addresses outside Wales screened out automatically. After additional drop-off at later stages, **201 participants** completed all three stages of the experiment.

The experiment proceeded through the following participant flow:

1. **Input postcode**
2. **View the explainer video**
3. **Complete multiple-choice comprehension questions**
4. **Verify cadastral data for their home parcel**
5. **Complete the valuation-method choice task**

This flow corresponds to the three stages described below. Each stage was designed to ensure that participants' preferences over valuation methods were as informed, consistent, and transparent as possible within the scope of a short online interaction.

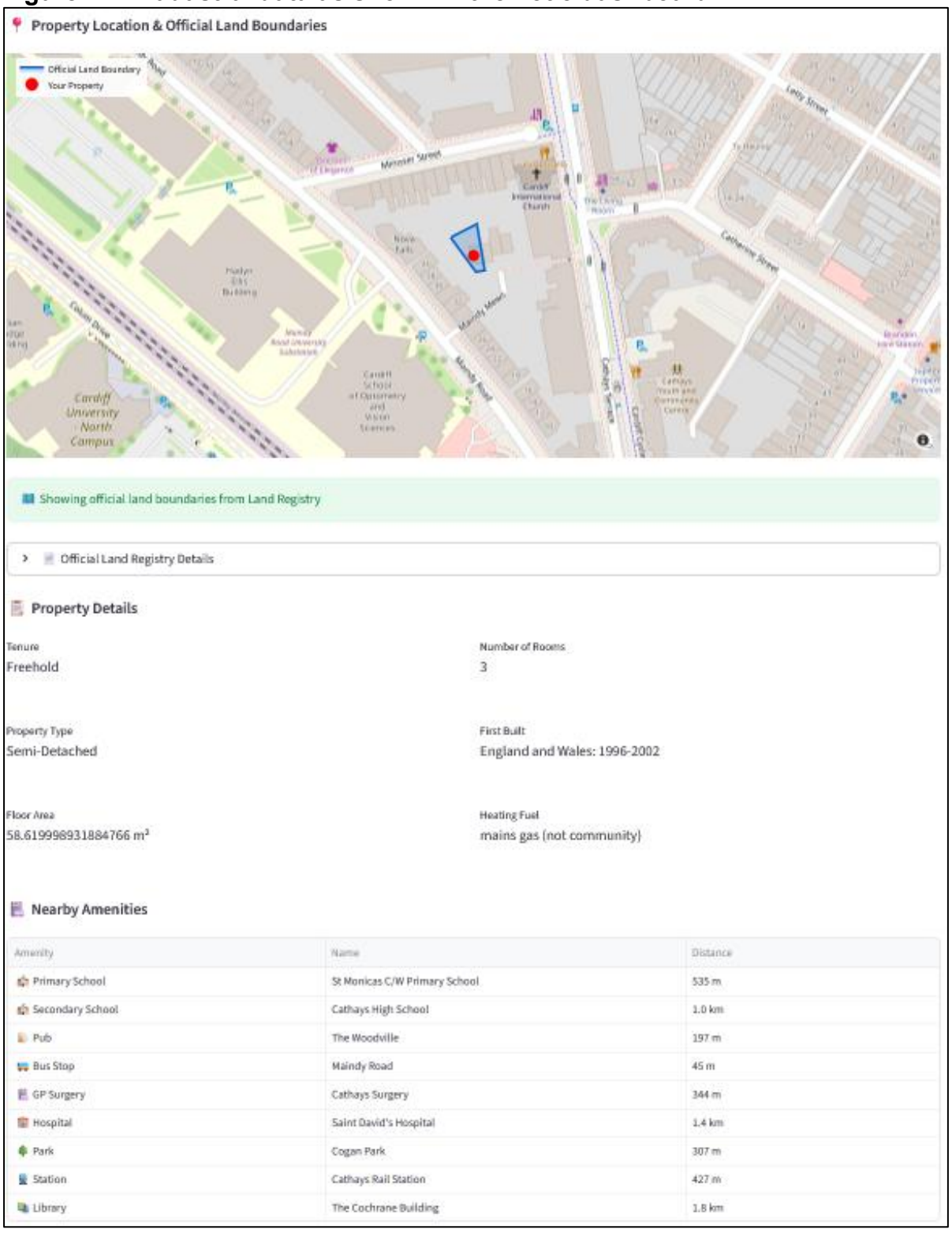
## STAGE 1: EDUCATION AND COMPREHENSION

In Stage 1, participants were shown a short explanatory animation introducing the concept of land valuation and the key distinctions between different modelling approaches. The purpose of this stage was not to train participants in technical modelling concepts, but to ensure that their eventual preferences were not based on misconceptions or confusion about what the competing valuation methods attempted to do.

Immediately after watching the video, participants completed a set of four comprehension questions. These questions tested only basic understanding of the video content and did not require any specialist knowledge.

This stage ensured that all subsequent decisions were made under comparable levels of understanding. Comprehension scores were recorded for later analysis but were not used to filter participants; rather, they allowed us to test whether comprehension predicted model preference (as reported in the findings section).

**Figure 2.1: Cadastral data as shown in the Lot 5 dashboard**




## STAGE 2: CADASTRAL DATA CONFIRMATION

Stage 2 provided participants with an interactive visualisation of the cadastral parcel we had linked to their address or postcode. Using the dashboard, participants were

shown the INSPIRE index polygon that we had matched to their address, and other data fields from the National Land Parcel Database including GIS derived spatial attributes and structure attributes from the EPC dataset. Figure 2.1 shows an example of the data presented.

Participants were asked to confirm whether these data appeared accurate and to rate the accuracy on a 1–10 scale, with an optional text box for comments. This stage served two purposes. First, it increased transparency by allowing participants to see and interrogate the data that underpinned the valuation models. Second, it allowed us to measure how confidence in data quality related to subsequent method choices, on the hypothesis that preferences might be conditional on beliefs about the accuracy of underlying information.

This data-confirmation component also enforced alignment with WRA principles by ensuring that people saw, checked, and understood the information used to generate valuations of their own land.

### STAGE 3: METHOD SELECTION UNDER INCENTIVES

After completing Stages 1 and 2, participants advanced to the core experimental task in Stage 3. They were presented with concise, parallel descriptions of the three land valuation methods developed in this project:

1. **Hedonic pricing** (Lot 1)
2. **Machine learning** (Lot 2)
3. **Equation-based method** (Lot 3)

For each method, participants could expand an information box to view a short explanation in consistent language, with emphasis on transparency, complexity, data requirements, and how each method relates to market values.

Participants were then required to choose *one* of the three methods. Crucially, this choice was **incentive-compatible**: each participant's final payout depended on the land value generated for their parcel by the selected method.

Participants were randomly assigned to one of two incentive treatments:

- **Low-valuation incentive treatment**
  Participants earned **£10 minus 0.001% of the land value** produced by the method they chose.
- **High-valuation incentive treatment**
  Participants earned **£5 plus 0.001% of the land value** produced by the method they chose.

This design created opposing financial motives across the two groups. Participants in the Low treatment were financially rewarded for choosing methods that returned lower land values, whereas participants in the High treatment were rewarded for

choosing methods that produced higher land values. Random assignment ensured that any systematic difference in method choice could be causally attributed to the incentive structure, not to underlying participant characteristics.

The value of 0.001% was chosen to keep payouts within the target range while ensuring that the incentives were salient. For a property with a land value of £100,000, the component linked to valuation would amount to £1.00.

Across the 201 participants who completed all three stages, average realised earnings were approximately **£8 per person**, although total project costs were higher once payments to screened-out participants and platform fees were included.

## DIGITAL PLATFORM AND TECHNICAL IMPLEMENTATION

The digital platform supporting the experiment was developed specifically for this project. While Stage 3 constitutes the behavioural experiment, Stages 1 and 2 rely heavily on the digital interface to deliver comprehension materials, present the interactive cadastral map, and pull in parcel-level data from the same datasets used across Lots 1–3.

The system was implemented in Python using a Streamlit-based architecture, with geospatial processing and rendering via GeoPandas, Shapely, and Plotly. Parcel boundaries from the INSPIRE Index Polygon dataset were retrieved and displayed dynamically once a participant entered a postcode and selected an address. Coordinate transformations and spatial indexing were applied to ensure that mapping outputs remained accurate while preserving application performance. All functionality – postcode screening, randomised treatment assignment, comprehension scoring, dashboard interaction, and method-choice logging – took place within a single integrated environment hosted at Bangor University. Prolific was used solely for recruitment and payment processing; no participant data were shared with Prolific.

The system is designed to scale: it can incorporate additional valuation methodologies, extend coverage to new geographies, and support broader engagements or public consultations. However, for the purposes of this project, it served to ensure that each step of the experimental protocol was delivered consistently and transparently to all participants.

## DATA

The modelling work undertaken in this project is underpinned by three comprehensive databases that we have constructed from multiple sources. These datasets form the foundation for all subsequent analysis across Lots 1, 2, 3, and 5.

The first database is a Wales-wide Transactions Database, representing as many property transactions as possible. To build this, we merged data from almost all of the identified sources, including HM Land Registry Price Paid data, INSPIRE Index Polygons, Energy Performance Certificate data, OS MasterMap Topography Layer, and socio-economic indicators such as the Welsh Index of Multiple Deprivation. The resulting dataset combines transaction prices with a rich set of property attributes, both structure-related and land-related, for each observation. After extensive cleaning and integration, this database contains just over one million transactions, providing the basis for estimating the various parameters required in Lots 1, 2 and 3.

The second database is a National Land Parcel Database. This is based on a subset of HM Land Registry INSPIRE Index Polygons within Wales. The final set of 1.4 million polygons selected were augmented with detailed land-related attributes derived from geospatial analysis. This database was designed to provide the backbone for the delivery of the Lot 5 dashboard, but also provides a useful basis for assessing and illustrating the results of the other three lots.

The third database is the LSOA Land Parcel Database covering the nine LSOAs specified in the project brief. This dataset provides maximum coverage of the land area within these nine LSOAs. This was enabled by augmenting HM Land Registry INSPIRE Index Polygons with polygons from the OS MasterMap Topographical Layer. Each of the 20,480 polygons in the dataset was then combined with detailed land-related attributes derived from geospatial analysis. This database was designed to support calculations of land valuations in Lots 1 and 2 and to enable formula-based approaches to valuation. While its geographic scope is narrower than the other two datasets, it provides the granularity required for parcel-level modelling within the specified LSOAs and for visual outputs such as maps.

Together, these three databases represent a substantial data engineering effort and provide the analytical backbone for all valuation methodologies tested in this project.

This section begins with a description of the primary data sources that underpin the modelling work. We also note data that we were unable to access or integrate into our main datasets. We then outline the processes used to compile and integrate these sources into three comprehensive databases: the Transactions Database, the National Land Parcel Dataset and the LSOA Land Parcel Database. Following this, we provide an overview of the characteristics of these three databases, including their scope and intended use within the project. The section concludes with a

discussion of overall limitations and a summary of the quality assurance measures applied throughout the data preparation process.

## PRIMARY DATA SOURCES

The project draws on a range of datasets that provide property values, land-related attributes, structure-related characteristics, and contextual indicators. Table 2.1 summarises the main variables and their sources. These datasets collectively underpin the two core databases constructed for the project:

- The HM Land Registry Price-Paid Dataset provides the backbone for the Transactions Dataset.
- HM Land Registry INSPIRE Index Polygons provide the basis for delineating parcels of land in the National Land Parcel Database.
- For the LSOA Land Parcel Database, HM Land Registry INSPIRE Index Polygons are combined with OS MasterMap Topography Layer polygons to delineate parcels of land.

### HM LAND RESISTRY PRICE PAID DATASET (PPD)

**Description:** Contains records of residential property transactions in England and Wales, including sale price, transaction date, property type, tenure, and address details.

**Scale and time-period:** The PPD dataset covers transactions since 1995. This amounted to 1,457,730 transactions in Wales at the time of download in September 2025.

**Relevance:** Essential for estimating property values and decomposing them into land and structure components for hedonic modelling.

**Usage in Project:** Used extensively to build the transaction-based dataset for Wales-wide modelling. Provided the dependent variable (property price) and tenure information.

**Limitations:**

- Does not include all non-domestic transactions.
- Limited coverage of unimproved land transactions (very few records).
- Requires cleaning for duplicates and address standardization.
- No geographic data.

**Table 2.1: Overview of data collected**

| Usage / Variable | Data sources |
|---|---|
| Property values ($H_{i,t}$) | HM Land Registry Price-Paid Dataset<br>HM Land Registry Commercial and Corporate Ownership Dataset |
| **Land-related attributes ($x^L_{i,t}$):** | |
| Plot size | Derived via GIS analysis from:<br>HM Land Registry INSPIRE Index Polygons<br>OS MasterMap Topography Layer |
| Land parcel delineation | HM Land Registry INSPIRE Index Polygons |
| Building footprint area | OS MasterMap Topography Layer |
| Tenure | HM Land Registry Price-Paid Dataset and Commercial Ownership Dataset |
| Land conditions | OS MasterMap Topography Layer |
| Distance to various amenities e.g. schools, parks, hospitals, bus stops, train stations, leisure facilities, retail hubs. | Derived via GIS analysis from:<br>DataMapWales, NaPTAN, OpenStreetMap, OS FEMA retail areas |
| Quality of amenities - schools | 'My Local School' |
| Environmental factors:<br>- Demographic e.g. population density<br>- Economic e.g. employment, income<br>- Social e.g. crime rates<br>- Natural e.g. particulate pollution, flood risk | WIMD dataset<br>NRW flood risk data |
| **Structure-related attributes ($x^S_{i,t}$):** | |
| Building characteristics e.g floor area, room count, property type, construction date, fuel type | Energy Performance Certificate data |
| **Matching properties across datasets** | |
| Addresses, UPRN | OS AddressBase |

### HM LAND REGISTRY COMMERCIAL AND CORPORATE OWNERSHIP DATA (CCOD)

**Description:** Dataset containing details of commercial and corporate property ownership in England and Wales, including title numbers, ownership type, and property addresses.

**Relevance:** Potentially useful for identifying property transactions and ownership patterns beyond those captured in the PPD.

**Usage in Project:** Employed to detect additional transactions and ownership records that were not present in PPD, improving completeness of property market analysis.

**Limitations:**

- Does not include residential properties.
- No geographic data.
- Address data require cleaning and matching.
- In practice, we found no additional transactions of relevance, beyond those in the PPD database.

### ORDNANCE SURVEY ADDRESSBASE

**Description:** A national address dataset from Ordnance Survey linking postal addresses to Unique Property Reference Numbers (UPRNs) and geographic coordinates. Within this project, the dataset was accessed as a Wales-only licensed extract, focusing on addressable delivery points for property-level analysis.

**Scale and time-period:**

- Spatial Scale: Wales-wide coverage at property-level resolution
- Geographic Extent: Limited to Wales under project-specific licensing agreement
- Dataset Components Used: Delivery Point Address (DPA) layer
- Record Volume (Wales only): DPA: 1,603,392 delivery point address records and BLPU: 2,162,637 records
- Licensing Duration: 3 months initial licence: 3 months extended by 9 months in December 2025

**Relevance:** Critical for standardizing address data and enabling spatial analysis by providing consistent identifiers and location information.

**Usage in Project:** Used to match addresses from HM Land Registry datasets to UPRNs and assign accurate geographic coordinates for mapping and analysis. Also used for UPRN-based joins across EPC, OS MasterMap, and amenity datasets.

**Limitations:**

- **Licensing restrictions:** Access to AddressBase is controlled under Ordnance Survey licensing, limiting redistribution and requiring authorised use within the project.
- **DPA coverage limitation:** The DPA dataset only includes postal delivery points, meaning non-addressable properties (e.g., some public buildings or land parcels) are excluded.
- **BLPU usability constraint:** Although more comprehensive, the BLPU layer includes non-addressable features and lacks a clear mechanism to reliably isolate only addressable properties at scale.
- **Address matching complexity:** Variations in address formats between datasets (e.g., PPD vs DPA) make matching non-trivial and require preprocessing or fuzzy matching techniques.
- **Temporal update lag:** Newly built or recently modified properties may not be immediately reflected in the dataset, leading to potential gaps.
- **UPRN relationship complexity:** There can be one-to-many or many-to-one relationships between addresses and UPRNs, especially for subdivided or multi-occupancy properties.

### HM LAND REGISTRY INSPIRE INDEX POLYGONS

**Description:** A geospatial dataset from HM Land Registry providing polygon boundaries representing registered freehold land parcels in England and Wales. The dataset enables spatial representation of title extents but does not include all property types.

**Scale:** Over 1.6 million land parcels in Wales.

**Relevance:** Critical for linking property transactions to spatial attributes and for constructing parcel-level datasets for the nine LSOAs.

**Usage in Project:** Used to derive plot size and spatial relationships (e.g., proximity to amenities) via GIS analysis.

**Limitations:**

- Incomplete coverage: Does not include unregistered land, leading to gaps in spatial representation.
- Leasehold exclusion: No polygons for leasehold properties, including flats, apartments, and maisonettes.
- Polygon integrity issues: Presence of overlapping polygons, duplicate geometries, and containment inconsistencies (e.g., nested or misaligned parcels).
- Data consistency challenges: Requires significant preprocessing and validation before use in analytical workflows.

- Complex integration requirements: Spatial joins with datasets such as OS MasterMap and AddressBase require careful handling to avoid mismatches and double counting.

### ENERGY PERFORMANCE CERTIFICATE (EPC) DATA

**Description:** EPCs were introduced in 2007 as a requirement for property sales. This data provides building-level information on floor area, property type, construction date, energy efficiency ratings, and various other structure characteristics.

**Scale and time-period:**

- Spatial Scale: Wales-wide coverage at building-level resolution.
- Record Volume: 1,571,222 EPC records (including multiple lodgements per property).
- Temporal Coverage: Data spans from 2000 onwards, with the majority of records from post-2007 (introduction of EPC regulations).
- Data Snapshots Used: Initial dataset in September 2025 and Updated dataset December 2025.
- Update Frequency: Periodically updated (typically monthly) by the data provider.

**Relevance:** Supplies structure-related attributes for hedonic modelling and machine learning models.

**Usage in Project:** Used to capture building characteristics; integrated with transaction data.

**Limitations:**

- Temporal coverage bias: EPC data is largely unavailable for properties that have not transacted since 2007, with very limited records transactions before this period and no records prior to 2000.
- Multiple lodgements and duplication: The dataset contains 1,571,222 records for Wales, including multiple EPC lodgements per property, requiring deduplication and selection logic (e.g., most recent record).
- Missing UPRNs: Approximately 68,715 records (~4.37%) have no UPRN, limiting direct linkage to spatial and transactional datasets.
- Matching complexity with Land Registry data: Accurate integration requires careful UPRN-based joins, supplemented by address construction and distance-based validation, particularly for flats, apartments, and multi-occupancy buildings.
- Structural inconsistency across property types: Domestic, non-domestic, and public buildings have different schemas and attribute availability, requiring harmonisation before analysis.

- Data quality issues: Presence of missing, inconsistent, or inaccurately recorded attributes (e.g., floor area, build year, energy ratings) introduces uncertainty into modelling.

### OS MASTERMAP TOPOGRAPHY LAYER

**Description:** High-resolution geospatial dataset detailing building footprints, land parcels, and physical features.

**Scale and time-period:** The dataset was obtained from EDINA in October 2025 as a Wales-wide extract. Due to the large volume of the full dataset (over 23 million polygon features), only a subset of relevant themes was extracted to optimise processing and analysis. The themes retained for the project include Buildings, Land, Buildings/Structures, Structures, Land/Structures, and Land/Water, focusing on features necessary for land characterisation and building footprint analysis.

**Relevance:** Enables accurate calculation of plot size and spatial positioning for both datasets.

**Usage in Project:** Used for GIS-derived measures such as parcel area, building footprint, land-to-building ratio, other land characteristics, and proximity to amenities.

**Limitations:**

- Licensing restrictions prevent sharing raw data externally.
- Processing requires specialist GIS tools and significant computational resources.
- Does not contain UPRNs, requiring integration with AddressBase or INSPIRE polygons for property-level analysis.
- Features represent physical geography rather than legal/property boundaries, requiring interpretation when deriving parcel-level metrics.

### WELSH INDEX OF MULTIPLE DEPRIVATION (WIMD) DATASET (2019)

**Description:** Provides small-area indicators of deprivation across domains such as income, employment, health, and access to services.

**Scale and time-period:** The dataset is available at Lower Super Output Area (LSOA) level across Wales and is based on the 2019 release of WIMD. Although a newer version was released in December 2025, this project utilises the 2019 dataset to maintain consistency with the project timeline and other aligned datasets. The data was originally structured using LSOA 2011 boundaries and subsequently realigned to LSOA 2021 geographies for integration within the project.

**Relevance:** Used to capture socio-economic context for land valuation models and to test sensitivity of valuations to local conditions.

**Usage in Project:** Incorporated as area-level covariates in modelling and for descriptive analysis of LSOAs.

**Limitations:**

- Aggregated at LSOA level; cannot capture intra-area variation.
- Updates infrequent. The latest version of WIMD was released in December 2025, after this project had commenced. We have, therefore, used the previous (2019) version of the data.
- The dataset reflects 2019 conditions and may not represent more recent socio-economic changes, particularly given the availability of a newer 2025 release.

### NATURAL RESOURCES WALES (NRW) FLOOD RISK DATA

**Description:** Spatial dataset showing flood risk zones for rivers, sea, and surface water across Wales.

**Relevance:** Land at risk of flooding is likely to have lower development value than land which is safe from floods, particularly as climate change increases the frequency and severity of flooding events.

**Usage in Project:** A single overall flood-risk variable was derived by taking the maximum of the three discrete flood-risk types at the point location of the UPRN for each transaction or property. Where no unique UPRN was identified for a property, the centroid of the polygon was used.

**Limitations:** May not reflect future changes in land use or climate.

### 'MY LOCAL SCHOOL' DATA

**Description:** Locations and attributes of schools in Wales, including type, performance, and demographics. Welsh Government data.

**Relevance:** There is a clear consensus in the literature that the quality of local schools is capitalised into land values (see Gibbons and Machin, 2008, for a summary).

**Usage in Project:** We use Key Stage 4 results as a measure of secondary school quality and attendance figures as a proxy for the quality of primary schools. School locations were used to calculate distances from properties, and to identify the names of local schools in the Lot 5 dashboard.

**Limitations:**

- The data does not include a direct measure of primary school quality.
- Some missing data for Key Stage 4 results and attendance.

### OPEN STREET MAP

**Description:** Crowd-sourced geospatial data covering roads, buildings, and points of interest. OpenStreetMap provides a broad set of amenity features including parks, hospitals, colleges and other public or recreational sites.

**Relevance:** The material adds coverage in areas where official data are not available. These amenities support accessibility analysis and offer contextual information for the Lot 5 interface.

**Usage in Project:** OSM features were filtered by functional class and linked to parcels through nearest-neighbour searches. Distances to amenities such as colleges, parks and hospitals form part of the spatial variables used in both modelling and experimental work. Polygons were converted to centroids where required.

**Limitations:**

- Coverage and naming conventions vary.
- Some amenities were absent in sparsely mapped areas.
- Classifications reflect community tagging rather than formal categories.

### DATA MAP WALES GP SURGERIES

**Description:** Locations of general practice surgeries in Wales, including coordinates, names and addresses.

**Relevance:** Essential for assessing healthcare accessibility and service provision.

**Usage in Project:** Each parcel was linked to its nearest surgery through a nearest-neighbour search. Distances and associated identifiers were retained as part of the spatial variables used in modelling and in the presentation layer of Lot 5.

**Limitations:** Coverage depends on the completeness of the national dataset.

### NAPTAN BUS STOPS AND RAILWAY STATIONS

**Description:** Dataset of public transport nodes including bus stops and railway stations.

**Relevance:** Transport access forms an important part of local infrastructure and is often associated with land value patterns (for example, Wang et al, 2015, conclude that the number of bus stops within walking distance of a property is positively associated with the property's observed sale price). These indicators allow the modelling framework to reflect basic accessibility conditions.

**Usage in Project:** Distances to the nearest bus stop and railway station were calculated for each parcel using spatial nearest-neighbour methods. The identifiers

and stop names were carried into the dataset for descriptive and analytical work, including the Lot 5 interface.

**Limitations:**

- Some stop records contain limited descriptive information.
- Distances represent straight-line measures rather than network travel.
- No data on quality of transport nodes, for example frequency of service.

### ORDNANCE SURVEY FEMA RETAIL AREAS

**Description:** Spatial boundaries of retail centres classified by size and function.

**Relevance:** Retail access forms part of the general service environment and can influence local preferences. These measures help describe the commercial setting of each parcel and support both modelling and participant interpretation in Lot 5.

**Usage in Project:** Distances to each retail category were computed using a nearest-neighbour search across the three cluster types. Centroids of cluster hulls were used for consistent spatial comparison. The classification of the closest retail area was retained as a simple categorical field.

**Limitations:**

- The centroid approach cannot reflect the full shape of retail areas.
- Some smaller retail clusters fall outside the FEMA classification.

## DATA NOT COLLECTED

### PLANNING DATA

Wales operates a plan-led, consent-based system rather than a zoning regime. As a devolved nation, Wales sets its planning policy through national frameworks such as *Planning Policy Wales* (Welsh Government 2024) and *Future Wales* (Welsh Government 2021), with Local Development Plans (LDPs) prepared by local authorities. These LDPs provide policy guidance rather than fixed, parcel-level zoning designations, unlike many other jurisdictions (for example in the USA) where zoning maps classify each parcel into use categories such as residential, commercial, or industrial.

We attempted to source LDP spatial data for all 22 Welsh local authorities to include planning context in our models, but encountered several limitations:

- Most local authorities only offered policy maps in PDF format, without accompanying GIS data.
- Spatial datasets were available for a few, such as Powys (publicly accessible) and Gwynedd (obtained via request).

- Among the GIS datasets obtained, there was no uniformity in LDP implementation: each local authority categorised land uses differently.

Given these issues and the absence of a standardised, nationally available parcel-level planning classification, we did not integrate planning designations into our databases. As a result, the only proxy for planning context in our data is the existing building footprint, derived from OS MasterMap. We recognise that this represents a significant gap in capturing planning permissions and potential land use changes in Wales.

### QUALITY OF AMENITIES

While our databases include numerous measures of distance to amenities, we were unable to incorporate robust indicators of amenity quality beyond schools. We prioritised schools because the literature consistently identifies school quality as a key driver of land value, and relevant data was relatively accessible and straightforward to process. For other amenities, quality measures were either unavailable in a usable format or would have required substantial additional effort beyond the scope of this project.

For example, quality proxies for transport nodes, such as frequency of services at train stations or bus stops, exist but would have been time-consuming to collect and standardise. For other amenities, such as playgrounds, parks, and GP practices, it is unclear what constitutes an appropriate quality metric. We are aware of inconsistencies in the data we do have; for instance, the “hospital” category spans everything from major tertiary centres like University Hospital Wales to small cottage hospitals, yet our dataset treats them uniformly. Similarly, we did not include road network accessibility at all, as defining and operationalising a meaningful measure (e.g., proximity to major routes versus local roads) would require additional methodological development.

These gaps represent important limitations. While our approach captures physical accessibility, it does not fully reflect service quality or functional capacity, which are likely to influence land values. Addressing these issues would require more time and resources than were available for this project.

### HIGH-RESOLUTION ENVIRONMENTAL DATA

Although our databases incorporate a wide range of environmental and infrastructure indicators, primarily from the Welsh Index of Multiple Deprivation (WIMD) dataset, these measures are aggregated at the LSOA level. While suitable for capturing broad contextual effects, this level of granularity limits precision in modelling parcel-level land values. More detailed data exists for some variables, but was not accessible or feasible to process within the scope of this project.

A key example is broadband connectivity. Property-level data on broadband availability and speed is available, but we were unable to obtain this in a GIS-ready format suitable for integration. Similar constraints likely apply to other environmental factors, such as air quality or infrastructure accessibility, where more granular datasets may exist but would require significant effort to source, clean, and standardise. Given time and resource limitations, we relied on LSOA-level indicators as a practical compromise, recognising that this introduces an important limitation in the spatial precision of our models.

### HM LAND REGISTRY NATIONAL POLYGON DATASET

HM Land Registry maintains a comprehensive National Polygon Dataset that provides detailed cadastral boundaries and associated title information for all registered land parcels in Wales (and England). Unlike the free INSPIRE Index Polygons, which offer only basic geometry and title references, the National Polygon Dataset includes enriched attributes such as tenure details and ownership metadata. Access to this dataset requires a commercial licence, with costs of approximately £25,000 per year. Due to budget constraints, we were unable to incorporate this dataset into our analysis. Its absence limits the completeness of parcel-level modelling, particularly for parcels where INSPIRE coverage is incomplete or where additional ownership attributes would have been valuable.

## DATA COMPILATION AND INTEGRATION

The construction of our analytical datasets required extensive data engineering to merge multiple sources into coherent, model-ready structures.

### COMPILATION OF THE TRANSACTIONS DATABASE

The Transactions Database was designed to support estimation of parameters in the various models developed in Lots 1, 2 and 3. It combines transaction prices with a rich set of structure-related and land-related attributes for each property. The compilation process began with 1,457,730 Wales-based transactions from the HM Land Registry Price Paid Dataset (PPD), relating to the period 1995 to July 2025.

The most significant challenge was that the PPD, which forms the backbone of the Transactions Database, does not include Unique Property Reference Numbers (UPRNs) or spatial geometry. To overcome this, we used the OS AddressBase database to match UPRNs to PPD addresses, a process complicated by variations in address syntax and formatting.

Once UPRNs were established, this allowed for the integration of EPC data.

LSOAs were identified via postcodes, allowing the integration of the WIMD dataset.

Finally, transaction prices were inflation-adjusted using House Price Index data to create a scaled price variable for modelling.

### INDEXING OF PRICE PAID

As property prices change substantially over time, before modelling it is important to standardise these to have a consistent timeframe across the whole model training dataset. Our preferred approach to accomplishing this was to scale all prices to their 2025 value. Using UK House Price Index data by property type (e.g. flats, terraced, detached), the yearly house price index for each of the 22 Welsh local authorities were calculated. The index per region and property type at the time of sale was then compared to the 2025 index, and the price was scaled accordingly, such that each property was given a 2025 price. Where property type was unknown or non-residential, the overall house price index for the local authority was used to estimate the scaled price.

### INTEGRATION AND DERIVATION OF SPATIAL VARIABLES

Throughout the analysis, a consistent spatial framework was used with all calculated distances and geocoding done under the British National Grid coordinate system. This helped to avoid mismatches especially when working with data of mixed reference systems and to ensure that the derived variables are comparable across the same scales.

Distances to amenities, including schools, health facilities, transport nodes and retail clusters, were calculated using nearest-neighbour methods which were applied to point or centroid features. School quality measures were also linked to the dataset using standardised school codes. Retail clusters followed the FEMA classification, differentiating between primary high street locations and other forms of commercial concentration in other to ascertain the nearest retail centre to property locations.

Information on Topography was sourced from the OS MasterMap Topographic Area layer. Land surface types were identified by filtering keywords and analysing spatial intersections using a 0.5m buffer around each parcel point. This produced results of parcels with slopes, scrubs, rocky areas, and other land surface types that could influence land value, use and development potential. The analysis identified that a single parcel could fall into more than one topographic type.

Observations were linked to Land Registry INSPIRE index polygons and OS MasterMap polygons via geometry, allowing for the calculation of plot size, the footprint of the main building and the footprint of other buildings within the land parcel.

## CLEANING OF OBSERVATIONS

Transactions with identical price paid, date of transaction and address were removed. Transactions for which addresses could not be matched to a UPRN were also removed. Finally, a cleaning process to identify spurious observations was carried out. These related to potential bundled sales. We identified sets of transactions on the same day for various properties within the same local authority which has the same price paid. Where the price paid per square metre of floor area was more than 50% higher than the median for that property type in the local authority, we assumed that there was a high probability that these represented a bundled sale where all transactions had been allocated the aggregate bundle price. These suspected anomalies were removed from the database.

Following these processes, the final Transactions Database consists of 1,284,218 observations. Table 2.2 summarises the cleaning process.

**Table 2.2: Construction of the Transactions Database – number of observations**

| | |
|---|---|
| **Raw data: Wales-based transactions in the PPD dataset** | 1,457,730 |
| Deduplication – removal of transactions with identical price paid, date of transaction and address | - 56,921 |
| Transactions for which addresses could not be matched to a UPRN | - 93,501 |
| Removal of potential bundled sales | - 23,090 |
| **Final Transactions Dataset** | 1,284,218 |

## MISSING VALUES

Missing data was addressed systematically to maintain consistency and minimise bias. For numeric contextual variables, such as socio-economic indicators from WIMD and environmental measures, missing values were imputed using LSOA-level lookups or median imputation where appropriate. This ensured complete coverage for area-level attributes without introducing artificial variability. For property-specific features, such as floor area or EPC energy ratings, missing values were retained rather than imputed, as these characteristics vary uniquely by property and cannot be reliably inferred from neighbouring records. Boolean land-surface flags derived from OS MasterMap were defaulted to FALSE where data was absent, to avoid introducing spurious classifications. For distance-to-amenity measures, missingness was resolved using KD-tree nearest-neighbour spatial imputation, copying values from the closest property with complete geometry. All transformations and imputations were documented in the dataset metadata, and intermediate layers were preserved for validation.

## COMPILATION OF THE NATIONAL LAND PARCEL DATABASE

The primary role of the National Land Parcel Database was to provide the data underlying the Lot 5 dashboard. Participants were recruited for Lot 5 from across Wales and so broad coverage was prioritised.

A total of 1.6 million land parcels were downloaded from the HM Land Registry INSPIRE Index Polygon datasets for each of Wales's 22 local authority areas. Many of these polygons are overlapping, providing a challenge in matching these with UPRNs and address data. The processing pipeline was designed to prioritise granularity, reproducibility, and geometric integrity, while avoiding sampling and other probabilistic methods. All source files were harmonised to a common coordinate reference system and subjected to basic geometry validation. Fewer than 0.2% of parcels were removed at this stage, indicating a high baseline level of data quality. Subsequent filtering focused on parcel size, with lower and upper bounds applied uniformly across the dataset. Large parcels above 5,000 square metres were excluded, while very small parcels were preserved to avoid systematic bias against fine-grained land units.

The result of this processing was to reduce the total number of polygons to 1,410,066 with a reduced overall coverage relative to the original dataset. The mean area of the land parcels in the dataset is 456.5 m² and the median is 254.5 m². This confirms genuine granularity, and suitability as a basis for Lot 5 visualisations. 90%of parcels fall below 1,000 square metres, with 14.34% under 100 square metres.

A full spatial conflict analysis was conducted on the filtered dataset using batched spatial indexing. No material overlaps or containment conflicts were identified at the specified thresholds. This result confirms that the size-based filtering did not introduce geometric inconsistency and that the retained parcels form a largely non-overlapping set.

In summary, the processing of land parcels for this database preserved national coverage, emphasised granularity, enforced strict geometric and size rules, and removed almost all overlaps and containments. The methodology was transparent, consistent and reproducible.

Spatial attributes were derived for all land parcels in the database by repeating the processes described above for the Transactions Database. WIMD data and EPC data (where available) were matched to the polygons following the processes described previously.

## COMPILATION OF THE LSOA LAND PARCEL DATABASE

The LSOA Land Parcel Database was developed to provide comprehensive coverage of all land within the nine specified LSOAs. Its primary objective was to

capture the characteristics of land parcels at a level of granularity sufficient for parcel-level land valuation modelling, but to preserve wide coverage. Geometric integrity was not prioritised.

**Table 2.3: Construction of the LSOA Land Parcel Database**

| Process Summary | Role | Number of polygons | Total Area (m2) | % of LSOA Coverage |
|---|---|---|---|---|
| Raw INSPIRE polygons clipped to 9 LSOAs; with duplication and containment/ overlap issues | Batch 1 input | 10,444 | N/A | N/A |
| Cleaned, deduplicated and overlap issues resolved to up to 90, enriched with UPRN, DPA attributes, area metrics, and cleaned geometry | Batch 1 output | 10,178 | 257,762,477 | 77.20% |
| Raw OSMM topographic areas polygons recruited to make up for areas missing on INSPIRE polygons. | Batch 2 input | 211,273 | N/A | N/A |
| Topographic polygons cleaned, version-filtered, geometry-validated, grouped by OSMM_fid and enriched with area metrics | Batch 2 output | 7,936 | 31,521,498 | 9.44% |
| Third batch of fragmented/ derived polygons from OSMM topographic areas polygons | Batch 3 input | 2,367 | N/A | N/A |
| Batch 3 polygons enriched, reconciled, LSOA-joined, UPRN-resolved | Batch 3 output | 2,367 | 39,823,404 | 11.93% |
| Union of enriched Outputs from Batches 1, 2, and 3; final deduplicated polygon layer | Combined Output | 20,480 | 329,107,379 | 98.57% |

The compilation process began with HM Land Registry INSPIRE Index Polygons, which offer delineations of registered land parcels. While INSPIRE provides a robust foundation for cadastral mapping, its coverage is limited to registered land, leaving gaps for parcels without recorded transactions. To address this, the OS MasterMap Topography Layer was integrated to ensure representation of unregistered land.

INSPIRE polygons frequently exhibit overlaps, particularly where multiple titles exist for the same physical area. To preserve granularity, a systematic hierarchy was

applied: smaller polygons were retained wherever possible, reflecting the principle that finer delineations typically correspond to more precise cadastral records.

OS MasterMap Topography Layer polygons also frequently exhibit overlaps, and include multiple versions of virtually identical polygons.

Table 2.3 summarises the process of integrating the INSPIRE and MasterMap polygons and removing multiple versions and overlaps.

Despite the challenges posed by the raw data, the resulting LSOA Land Parcel Database achieves near-complete coverage (98.6%) of the nine LSOAs and provides a rich set of parcel-level attributes for modelling.

The land attributes for each identified parcel of land were derived in the same way as for the Transactions Database. This included parcel area, building footprint area, land surface classifications, distances to amenities, and environmental indicators identical to those in the Transactions Database.

## OVERVIEW OF THE TRANSACTIONS DATABASE

The final Transactions Dataset consists of 1,284,218 observations of property transactions in Wales since 1995. The average transaction price in the database (scaled to 2025 prices) is £258,392.

These statistics are shown in Tables 2.4 and 2.5, along with the corresponding figures for each property type. As can be seen, the database is dominated by residential transactions (at least 86.5%). Detached, semi-detached and terraced houses are close to being equally represented, with flats/maisonettes far less common.

Figure 2.2 provides an overview of the price paid, scaled to 2025 values, in these transactions.

**Table 2.4: Transactions Database summary statistics, by Land Registry property type**

| Land Registry Property Type | Number of Transactions | Proportion of Transactions | Average Scaled Price |
|---|---|---|---|
| Detached | 355,173 | 27.70% | £372,469 |
| Semi-detached | 370,861 | 28.90% | £235,796 |
| Terraced | 446,565 | 34.80% | £187,361 |
| Flats/maisonettes | 89,092 | 6.90% | £171,093 |
| Other | 22,527 | 1.80% | £585,137 |
| **TOTAL** | **1,284,218** | **100.00%** | **£258,392** |

**Table 2.5: Transactions Database summary statistics, by EPC property type**

| EPC Property Type | Number of Transactions | Proportion of Transactions | Average Scaled Price |
|---|---|---|---|
| **Residential** | 1,105,650 | 86.10% | £250,560 |
| **Unknown / Not Stated** | 163,374 | 12.72% | £285,895 |
| **Retail & Commercial** | 7,037 | 0.55% | £400,692 |
| **Food/Drink & Hospitality** | 3,112 | 0.24% | £358,381 |
| **Offices** | 2,032 | 0.16% | £640,441 |
| **Residential Institutions** | 1,473 | 0.11% | £969,709 |
| **Industrial & Warehousing** | 713 | 0.06% | £1,331,090 |
| **Healthcare** | 392 | 0.03% | £529,075 |
| **Community / Education / Public Buildings** | 264 | 0.02% | £553,908 |
| **Assembly & Leisure** | 140 | 0.01% | £690,017 |
| **Other** | 16 | 0.00% | £759,474 |
| **Emergency Services** | 8 | 0.00% | £190,395 |
| **Utilities & Infrastructure** | 4 | 0.00% | £475,223 |
| **Transport** | 2 | 0.00% | £6,327,996 |
| **Mixed Residential / Commercial** | 1 | 0.00% | £145 |
| **TOTAL** | 1,284,218 | 100.00% | £258,392 |

Land parcel size was derived for each transaction by geospatial matching of addresses in the PPD to INSPIRE polygons, using DPA UPRNs from OS Address Base. The resulting average land parcel size per transaction in the Transactions Database was 27,995m$^2$. The unadjusted mean is driven by a long-tail distribution of exceptionally large parcels in the INSPIRE database.

The parcel area for flats/maisonettes was generally larger than for other property types. This reflects the difficulty in matching flats/maisonettes transactions with the relevant INSPIRE index polygon. In order to address this issue, we derived a variable, Adjusted Freehold Parcel Area, which divides the raw parcel data across the number of DPA UPRNs contained within the relevant INSPIRE polygon. Where

there is a single DPA UPRN (e.g. houses), then the Adjusted Freehold Parcel Area is identical to the original parcel area. Where there are multiple DPA UPRNs (e.g. blocks of flats), the raw parcel area is thus shared across the UPRNs.

**Figure 2.2: Histogram of scaled price paid in the Transactions Database**

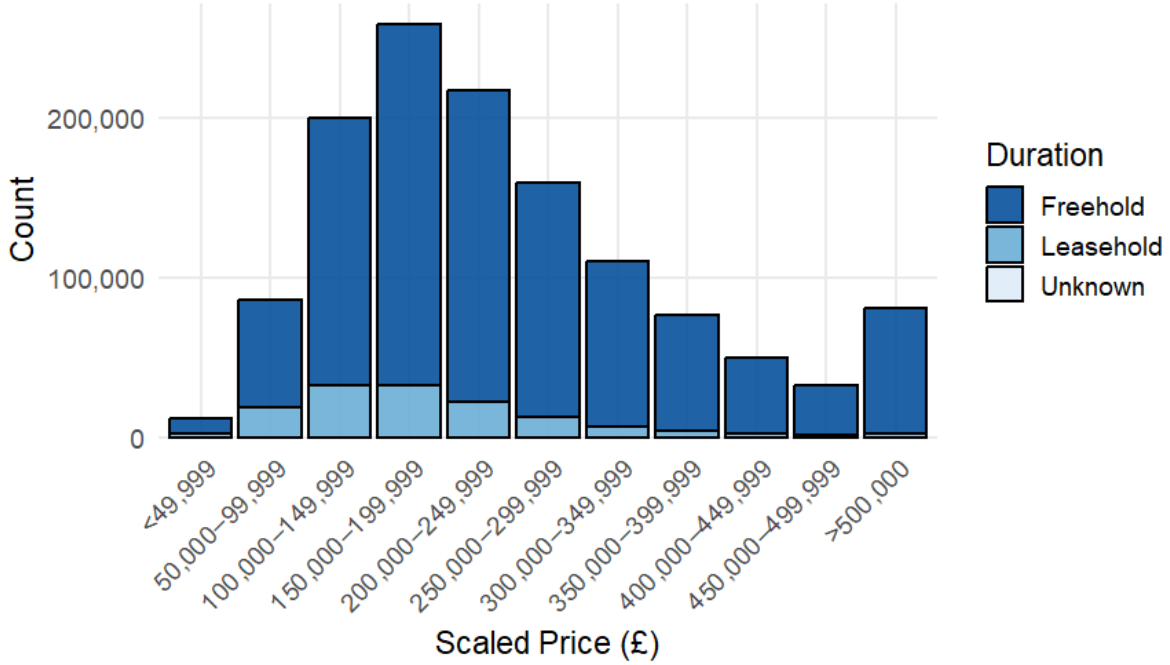



**Figure 2.3: Adjusted Freehold Parcel Area ($m^2$) by Land Registry Property Type (trimmed at 95th percentile)**

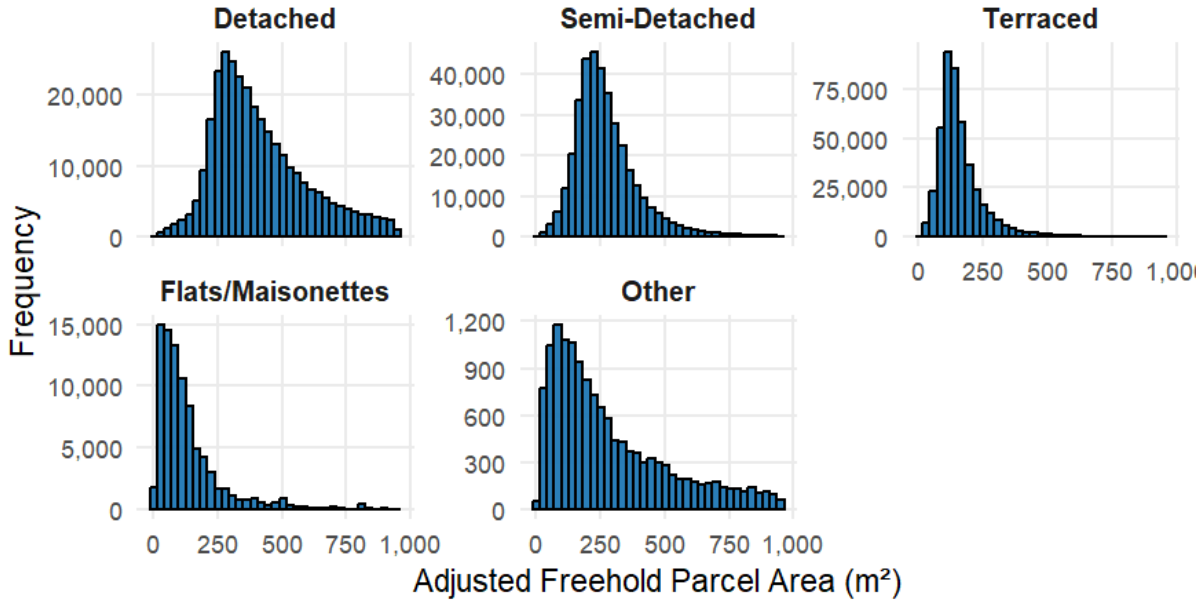


Figure 2.3 shows the distribution of Adjusted Freehold Parcel Area by property type, excluding the 5% of transactions with the largest land areas.

Table 2.6 shows the number of transactions in each local authority area, together with the average scaled price in that area. It also shows the number of transactions as a proportion of the local authority population (mid-year 2023 estimate). Every local authority has between 0.32 and 0.50 transactions per capita, with the lowest in Ceredigion and highest in Conwy and the Vale of Glamorgan. The highest average scaled price is over £387k in Monmouthshire. The lowest is just over £155k in Blaenau Gwent.

**Table 2.6: Transactions by local authority**

| Local Authority | Number of Transactions | Transactions per capita | Average Scaled Price |
|---|---|---|---|
| **Blaenau Gwent** | 24,524 | 0.36 | £155,388 |
| **Bridgend** | 61,694 | 0.42 | £251,892 |
| **Caerphilly** | 69,629 | 0.39 | £225,539 |
| **Cardiff** | 162,546 | 0.42 | £333,078 |
| **Carmarthenshire** | 71,328 | 0.38 | £232,249 |
| **Ceredigion** | 23,480 | 0.32 | £254,854 |
| **Conwy** | 57,475 | 0.50 | £257,772 |
| **Denbighshire** | 43,987 | 0.45 | £227,670 |
| **Flintshire** | 61,545 | 0.39 | £251,773 |
| **Gwynedd** | 43,187 | 0.36 | £236,090 |
| **Isle of Anglesey** | 26,401 | 0.38 | £264,520 |
| **Merthyr Tydfil** | 21,147 | 0.36 | £175,001 |
| **Monmouthshire** | 41,592 | 0.44 | £387,540 |
| **Neath Port Talbot** | 56,701 | 0.40 | £195,928 |
| **Newport** | 68,681 | 0.42 | £264,386 |
| **Pembrokeshire** | 49,259 | 0.39 | £260,447 |
| **Powys** | 45,111 | 0.34 | £268,121 |
| **Rhondda Cynon Taf** | 103,331 | 0.43 | £194,697 |
| **Swansea** | 102,576 | 0.42 | £249,208 |
| **Torfaen** | 34,426 | 0.37 | £228,626 |
| **Vale of Glamorgan** | 67,209 | 0.50 | £345,810 |
| **Wrexham** | 48,389 | 0.36 | £243,371 |
| **TOTAL WALES** | 1,284,218 | 0.41 | £258,392 |

Tables A1, A2, and A3 in the Appendix show the summary statistics for the numerical variables in the final Transactions Database. Tables A4, A5, and A6 show the corresponding summary statistics for the transactions which took place within the nine identified LSOAs.

## OVERVIEW OF THE NATIONAL LAND PARCEL DATABASE

The final National Land Parcel Database consists of 1,410,066 polygons, accounting for a total of 643km$^2$, or slightly more than 3% of the land area of Wales. The distribution of these polygons across local authorities is shown in Table 2.7. As can be seen, coverage is higher in urban areas and lower in rural areas. This reflects the data processing procedures, which prioritised granularity, as well as the nature of the underlying data.

**Table 2.7: Coverage of the National Land Parcel database by local authority**

| Local Authority | Area / km$^2$ | Number of polygons | Aggregate area of polygons / km$^2$ | Coverage | Mean polygon size / m$^2$ |
|---|---|---|---|---|---|
| **Blaenau Gwent** | 109 | 30,959 | 9.6 | 8.9% | 313.0 |
| **Bridgend** | 251 | 65,902 | 24.3 | 9.7% | 369.6 |
| **Caerphilly** | 277 | 81,012 | 27.1 | 9.8% | 334.9 |
| **Cardiff** | 141 | 121,483 | 37.2 | 26.4% | 306.6 |
| **Carmarthenshire** | 2,370 | 97,046 | 57.0 | 2.4% | 588.1 |
| **Ceredigion** | 1,785 | 37,147 | 26.4 | 1.5% | 713.1 |
| **Conwy** | 1,126 | 53,049 | 26.2 | 2.3% | 495.5 |
| **Denbighshire** | 837 | 44,421 | 22.8 | 2.7% | 514.7 |
| **Flintshire** | 440 | 68,220 | 32.8 | 7.5% | 482.1 |
| **Gwynedd** | 2,535 | 65,786 | 38.3 | 1.5% | 582.5 |
| **Isle of Anglesey** | 712 | 35,324 | 23.0 | 3.2% | 653.5 |
| **Merthyr Tydfil** | 111 | 26,805 | 8.4 | 7.7% | 316.9 |
| **Monmouthshire** | 849 | 45,149 | 27.8 | 3.3% | 615.9 |
| **Neath Port Talbot** | 441 | 67,926 | 26.9 | 6.1% | 396.7 |
| **Newport** | 190 | 63,838 | 22.4 | 11.8% | 351.0 |
| **Pembrokeshire** | 1,618 | 65,263 | 39.9 | 2.5% | 612.0 |
| **Powys** | 5,181 | 70,549 | 52.1 | 1.0% | 739.4 |
| **Rhondda Cynon Taf** | 424 | 117,629 | 34.6 | 8.2% | 294.9 |
| **Swansea** | 378 | 99,175 | 39.8 | 10.5% | 402.1 |
| **Torfaen** | 126 | 37,600 | 13.2 | 10.5% | 352.5 |
| **Vale of Glamorgan** | 331 | 60,075 | 24.8 | 7.5% | 413.1 |
| **Wrexham** | 504 | 53,436 | 25.2 | 5.0% | 473.1 |
| **NA (e.g. cross-boundary)** | N/A | 2,272 | 1.8 | N/A | 832.8 |
| **TOTAL** | **20,736** | **1,410,066** | **643.0** | **3.1%** | **456.0** |

## OVERVIEW OF THE LSOA LAND PARCEL DATABASE

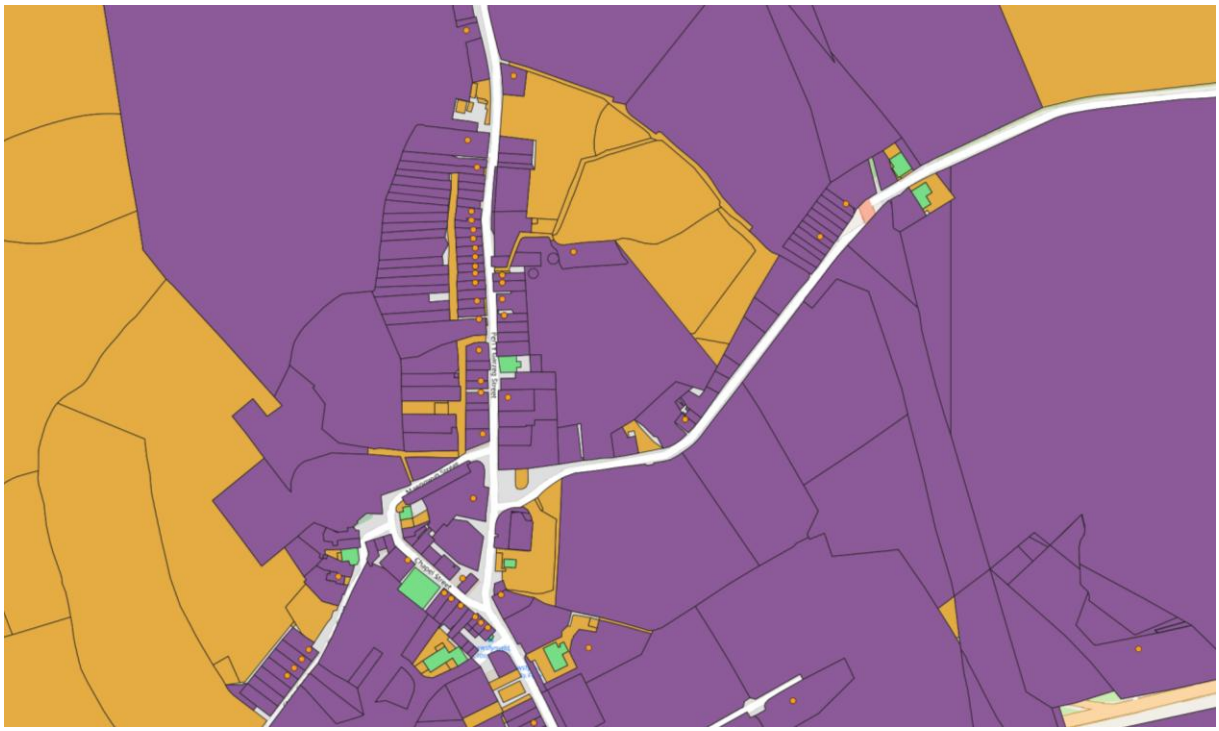
The final LSOA Land Parcel Database consists of 20,480 polygons, accounting for 98.6% of the land area within the nine identified LSOAs. The distribution of these polygons across the LSOAs is shown in Table 2.8. The lower coverage in Ceredigion 002D and Cardiff 032H reflect the higher proportion of land in these LSOAs accounted for by roads and rivers.

**Table 2.8: Coverage of the LSOA Land Parcel Database**

| LSOA code | Area of LSOA / $m^2$ | Number of polygons | Aggregate area of polygons / $m^2$ | Coverage |
|---|---|---|---|---|
| **Trawsfynydd** | 163,380,108 | 5,652 | 161,808,461 | 99.04% |
| **Broughton North East** | 9,211,005 | 1,639 | 8,982,571 | 97.52% |
| **Knighton 2** | 11,834,016 | 1,280 | 11,596,874 | 98.00% |
| **Aberystwyth Rheidol 1** | 162,251 | 616 | 121,588 | 74.94% |
| **Scleddau** | 71,108,643 | 4,105 | 69,958,385 | 98.38% |
| **Porthcawl East Central 2** | 792,619 | 1038 | 749,772 | 94.60% |
| **Rhigos** | 74,189,795 | 3,961 | 72,982,035 | 98.37% |
| **Usk 2** | 1,481,007 | 1,070 | 1,405,863 | 94.93% |
| **Cathays 12** | 1,731,469 | 1119 | 1,501,831 | 86.74% |
| **TOTAL for 9 LSOAs** | **333,890,913** | **20,480** | **329,107,379** | **98.57%** |

Figure 2.4 shows an extract from the LSOA Land Parcel Database, overlaid with point data representing observations from the Transactions Database. Purple polygons are derived from the HM Land Registry INSPIRE index dataset. Green coloured polygons are derived from the OS MasterMap dataset and have a one-to-one match with a UPRN. Mustard coloured polygons are derived from the OS MasterMap dataset but do not have a one-to-one match with a UPRN. Yellow points represent the UPRNs of properties that appear in the Transactions Database. As can be seen, the majority of areas with no polygons correspond to highways which, along with water-courses, were deliberately excluded from the recruitment of OS MasterMap polygons.

**Figure 2.4: Illustrative extract from the LSOA Land Parcel Database, centre of Trawsfynydd**



## SUMMARY

The three datasets described in this Section serve as the empirical foundation for the modelling results and behavioural evidence presented in Section 3. These datasets enable consistent comparisons across lots and ensure that differences in results reflect modelling choices rather than inconsistent data.

# 3. FINDINGS

## LOT 1: MARKET BASED STATISTICAL VALUATION

### MODEL FIT

The final model estimates land values using a hedonic regression in which the total transaction price, expressed in logarithmic form, is the outcome variable. During model development, a range of alternative specifications were tested, including models based on price per area. The log of total price was ultimately selected because it produced more stable and consistent results. The transformation improves statistical performance by reducing the influence of extreme values and modelling price variation in proportionate terms.

The logarithmic transformation reduces right skew in the distribution of transaction prices and compresses the scale of high-value observations, limiting the influence of extreme transactions. Differences in site size are explicitly accounted for through the inclusion of parcel area and related characteristics within the model specification, allowing price variation to be estimated in an economically meaningful way. Importantly, the model generates valuations at the parcel level rather than at the level of individual dwellings. This reflects the underlying structure of the data and ensures that estimated values are interpreted as relating to land parcels.

For transparency, the report also presents results from the alternative specification based on price per unit area so that the implications of each modelling approach can be directly compared.

The overall fit of the model is summarised using $R^2$, which measures the proportion of variation in logged transaction prices explained by the structural, locational, environmental, and socio-economic characteristics included in the model. Because the dependent variable is specified in logarithmic form, the $R^2$ reflects how well the model explains proportional differences in prices across parcels rather than absolute differences in cash values. $R^2$ takes a value between 0 and 1, where 0 indicates no explanatory power and 1 indicates perfect explanation of observed variation.

For the final hedonic regression model, the **$R^2$ is 0.441**. This indicates that approximately 44.1% of the variation in proportional land prices across parcels is explained by the variables included in the model. Land and property markets are inherently diverse and influenced by many factors that are not directly observed in the data, such as negotiation dynamics, site-specific features, and buyer expectations. As a result, $R^2$ values of this magnitude are common in parcel-level hedonic modelling.

In addition to $R^2$, the adjusted $R^2$ is also reported. This measure accounts for the number of explanatory variables included in the model and provides a more

conservative assessment of model fit. The adjusted $R^2$ for the final model is 0.440, indicating that explanatory power remains effectively unchanged after accounting for model complexity.

### MODEL FIT BY PROPERTY TYPE

Performance differs significantly by property type. The training dataset is heavily dominated by residential transactions, which account for the vast majority of observations. As a result, model performance is strongest for residential parcels.

For residential properties, the model achieves an **$R^2$ of 0.333**, indicating that approximately one third of observed variation in transaction prices within this category is explained by the characteristics included in the model.

In contrast, non-residential categories are represented by far fewer observations and encompass a much wider range of property types and site characteristics. This contributes to lower and more variable model fit across these groups. Retail, office, industrial, and public building categories each have relatively small sample sizes and show substantially lower explanatory power. In a small number of categories with very limited observations and substantial internal diversity, explanatory power is minimal. This reflects data constraints within those specific groups rather than instability in the overall modelling framework.

Results for non-residential categories should therefore be interpreted with appropriate caution, recognising the smaller evidence base available for those segments of the market. $R^2$ for each property type is reported in Table 3.1.1.

### MODEL FIT BY LOCAL AUTHORITY

Model performance also varies by local authority area. This reflects differences in market structure, price dispersion, and the composition of transactions across authorities.

$R^2$ values range from approximately 0.05 to 0.40 across local authorities. In some areas, particularly those with more consistent market structures or less price dispersion, a relatively higher proportion of variation in logged transaction prices is explained by the model. For example, **Rhondda Cynon Taf ($R^2$ = 0.40), Powys (0.34),** and the **Isle of Anglesey (0.29)** show comparatively stronger explanatory power.

In contrast, more urban and economically diverse areas such as **Cardiff ($R^2$ = 0.07)** and **Swansea (0.07)** exhibit lower explanatory power. This does not indicate model instability. Rather, it reflects greater internal diversity in property types, neighbourhood effects, and transaction dynamics, all of which increase price variation that cannot be fully captured through observable characteristics alone.

Overall, differences in model fit across authorities are primarily driven by underlying market complexity and structural diversity rather than changes in model specification.

The overall model reports a higher $R^2$ than models estimated separately by property type or local authority. This is expected and reflects how the data are structured rather than any weakness in the model. The pooled (overall) model explains variation both within and between locations and property types. Differences in average land values across locations and property categories account for a significant share of total price variation. When these structural differences are included in a single model, the proportion of variance explained increases. By contrast, models estimated within individual property types or local authorities remove these between-area and between-type differences. They therefore focus only on the smaller, more detailed variation within a single segment of the market. That remaining variation is naturally harder to explain, which leads to lower $R^2$ values. The lower $R^2$ in sub-segment models is therefore a normal statistical outcome and does not indicate weaker performance.

**Table 3.1.1: Model fit and performance by property type**

| Property type | Number of Transactions | RMSE (£) | $R^2$ |
|---|---|---|---|
| **Residential** | 1,105,650 | 147,807 | 0.333 |
| **Unknown or Not Stated** | 163,374 | 425,002 | 0.088 |
| **Retail and Commercial** | 7,037 | 1,845,303 | 0.022 |
| **Food/Drink and Hospitality** | 3,112 | 859,834 | 0.040 |
| **Offices** | 2,032 | 2,237,849 | 0.021 |
| **Residential Institutions** | 1,473 | 3,308,857 | 0.003 |
| **Industrial and Warehousing** | 713 | 5,237,765 | -0.020 |
| **Healthcare** | 392 | 1,584,691 | 0.030 |
| **Community, Education and Public Buildings** | 264 | 1,610,672 | -0.005 |
| **Assembly and Leisure** | 140 | 1,556,713 | 0.019 |
| **TOTAL** | 1,284,218 | 315,552 | 0.441 |

## KEY DRIVERS

In addition to model fit, it is important to consider the key economic drivers of predicted land values. The strongest driver in the model relates to parcel size, captured through the adjusted freehold parcel area variable. Larger parcels are associated with substantially higher total transaction values, reflecting the fundamental role of land scale in determining price. Property category is also a significant determinant of value. Transport-related uses are associated with

materially higher transaction values relative to the reference category, while mixed residential and commercial properties are associated with substantially lower values once size and location are held constant. Location effects are economically meaningful. Differences across local authorities remain statistically significant even after controlling for parcel size, property type and environmental characteristics. This indicates that underlying local market conditions continue to influence price formation. A range of environmental, accessibility and socio-economic indicators are also statistically significant. While these effects are smaller in magnitude than parcel size and property category, they collectively contribute to explaining spatial variation in land values across Wales.

**Table 3.1.2: Model fit and performance by local authority**

| Local authority | Number of Transactions | RMSE (£) | $R^2$ |
|---|---|---|---|
| **Blaenau Gwent** | 24,524 | 112,939 | 0.177 |
| **Bridgend** | 61,694 | 279,630 | 0.108 |
| **Caerphilly** | 69,629 | 329,485 | 0.076 |
| **Cardiff** | 162,546 | 519,079 | 0.072 |
| **Carmarthenshire** | 71,328 | 333,713 | 0.069 |
| **Ceredigion** | 23,480 | 140,710 | 0.165 |
| **Conwy** | 57,475 | 156,727 | 0.261 |
| **Denbighshire** | 43,987 | 163,802 | 0.234 |
| **Flintshire** | 61,545 | 309,578 | 0.072 |
| **Gwynedd** | 43,187 | 202,572 | 0.134 |
| **Isle of Anglesey** | 26,401 | 155,430 | 0.288 |
| **Merthyr Tydfil** | 21,147 | 222,161 | 0.082 |
| **Monmouthshire** | 41,592 | 597,229 | 0.053 |
| **Neath Port Talbot** | 56,701 | 166,612 | 0.169 |
| **Newport** | 68,681 | 253,011 | 0.175 |
| **Pembrokeshire** | 49,259 | 259,922 | 0.101 |
| **Powys** | 45,111 | 141,545 | 0.338 |
| **Rhondda Cynon Taf** | 103,331 | 108,421 | 0.400 |
| **Swansea** | 102,576 | 406,328 | 0.066 |
| **Torfaen** | 34,426 | 214,031 | 0.116 |
| **Vale of Glamorgan** | 67,209 | 256,503 | 0.250 |
| **Wrexham** | 48,389 | 195,078 | 0.223 |
| **TOTAL** | 1,284,218 | 315,552 | 0.441 |

## PREDICTIVE PERFORMANCE

While $R^2$ summarises explanatory power in logarithmic terms, predictive performance is assessed using the Root Mean Squared Error (RMSE). RMSE measures the typical difference between observed and predicted transaction prices in monetary terms. Because predictions are transformed back into total price in level space, RMSE is expressed directly in pounds and provides a practical measure of valuation accuracy.

For the full training dataset, the **RMSE is £315,552**. This represents the approximate average magnitude of prediction error across parcels included in the model. Because parcel values vary substantially in scale, absolute errors naturally increase for higher-value properties. The use of a logarithmic modelling framework ensures that predictions remain proportionate across the price distribution rather than being unduly influenced by extreme high-value transactions.

## PERFORMANCE BY PROPERTY TYPE

Predictive accuracy differs materially by property type, reflecting differences in transaction scale, price dispersion, and sample size across categories. Residential transactions dominate the dataset and exhibit an **RMSE of £147,807**. Given the scale of residential land values and the inherent variability of property markets, this level of error is consistent with parcel-level valuation modelling. Non-residential categories, by contrast, represent a much smaller share of observations and encompass a wide range of property types and transaction values. As a result, RMSE values are materially higher in absolute terms. For example, Retail and Commercial transactions show an RMSE of £1.85 million, Offices £2.24 million, and Industrial and Warehousing £5.24 million. These figures reflect the substantially larger average transaction values and wider dispersion typically observed in commercial property markets. In addition, several non-residential categories have very small sample sizes. In such cases, prediction error is influenced both by limited data availability and by structural diversity within each category. These patterns reflect underlying market characteristics rather than instability in the modelling framework.

## PERFORMANCE BY LOCAL AUTHORITY

Predictive accuracy also varies across local authorities. RMSE values reflect differences in overall price levels, market structure, and price dispersion within each authority. Across Wales, RMSE ranges from approximately £108,421 in Rhondda Cynon Taf to £597,229 in Monmouthshire. Authorities with higher overall price levels and greater dispersion in transaction values, such as **Cardiff (£519,079)** and **Swansea (£406,328)**, exhibit higher absolute prediction errors. This reflects the wider spread of transaction prices in these markets rather than instability in the modelling framework. Conversely, authorities with lower average transaction values or more compressed price distributions, such as **Rhondda Cynon Taf (£108,421)**,

**Blaenau Gwent (£112,939)**, and **Powys (£141,545)**, show lower absolute RMSE values. Because RMSE is measured in pounds, it naturally increases with the scale of prices. These differences therefore reflect variation in underlying market conditions rather than differences in model specification across authorities.

## ERROR DISTRIBUTION

To assess whether the model performs consistently across low-value and high-value parcels, predictive accuracy was examined across the price distribution. Parcels were grouped into ten equally sized groups (deciles) based on observed transaction value. Within each group, the median observed price, median predicted price, and median absolute and relative prediction errors were calculated.

Model performance is strongest in the middle of the price distribution. For deciles 3 to 7, the median relative error ranges between approximately 14% and 18%. In these segments, predicted median values closely track observed medians, indicating stable proportional performance.

At the lower and upper ends of the distribution, relative error increases. The lowest decile exhibits a median relative error of approximately 74%. This reflects the small absolute transaction values in this segment, where even modest cash differences translate into large percentage movements. The highest decile shows a median relative error of approximately 34%, reflecting greater structural diversity and price dispersion among higher-value parcels.

Importantly, error increases gradually rather than disproportionately across the upper deciles, indicating that the logarithmic specification maintains proportionate performance even for higher-value parcels. This pattern is consistent with empirical evidence from land and property valuation models in comparable markets.

## MODEL STABILITY AND ROBUSTNESS

To ensure that the estimated relationships reflect stable market patterns rather than being unduly influenced by particular transactions or segments of the data, the model was re-estimated on a large random subsample of 200,000 observations. The direction (sign) of each coefficient was then compared with the corresponding estimate in the full-sample model.

Across the subsample, approximately 89% of coefficients retained the same sign as in the full model. This high level of sign agreement indicates that the core relationships between characteristics and transaction prices are stable and not driven by a specific subset of observations. While minor variations are expected in any large-scale empirical model, particularly for variables with weaker explanatory power or smaller effective sample sizes, the overall consistency observed here provides reassurance that the model specification is robust.

## RESIDUAL BEHAVIOR

Residuals from the logarithmic specification were examined to assess whether the model tends to consistently over or under-predict transaction prices and to evaluate the adequacy of the chosen functional form. Across a large random sample of 300,000 observations, the mean residual in log space is 0.0005, indicating that the model does not display a consistent pattern of over or under-prediction. The median residual is similarly close to zero, and the distribution of residuals is broadly symmetric, with the 5th and 95th percentiles at -0.69 and 0.59 respectively.

The relationship between predicted values and residuals was also examined. The correlation between the two is effectively zero (-0.001), indicating that prediction errors do not systematically increase or decrease as predicted values rise. This provides reassurance that the model performs consistently across the fitted price range. Overall, the residual diagnostics provide reassurance that the model is statistically well-behaved and that prediction errors are not driven by flaws in the chosen specification.

## ALTERNATIVE OUTCOME VARIABLE

As part of the model development process, the hedonic regression was also estimated using the natural logarithm of price per unit area as the dependent variable. This approach models transaction values on an intensity basis rather than total parcel value.

This alternative outcome variable specification achieves an **$R^2$ of 0.314**, compared with 0.441 for the final selected model. When predictions are transformed back into total price in level terms, the corresponding RMSE is substantially higher than £315,552 observed under the preferred specification.

While the alternative specification produces broadly comparable explanatory performance, the log of total price model was preferred for three principal reasons. First, it aligns directly with parcel-level valuation, which is the focus of the analysis. Second, modelling total price avoids potential distortions introduced by dividing by area in cases where recorded parcel sizes are small or variable. Third, predictive diagnostics indicated more stable proportional performance across the upper end of the price distribution under the total price specification.

In addition, alternative functional forms were tested, including models estimated in levels and a Gamma-based specification. These did not produce superior overall performance or stability and were therefore not adopted.

## ECONOMIES OF SCALE IN LAND PRICING

It is important to recognise that land pricing exhibits scale effects. Smaller parcels typically transact at higher prices per unit area than larger parcels. This reflects a

range of economic factors, including greater liquidity, a wider pool of potential purchasers, and the flexibility associated with smaller sites.

As parcel size increases, price per unit area tends to decline. In the preferred specification, parcel area is included explicitly within the model, allowing scale effects to be captured empirically. However, it is recognised that such effects may not be fully linear or completely captured by observable characteristics alone.

By modelling total transaction price and including area as an explanatory variable, the relationship between size and value is estimated directly rather than mechanically imposed by dividing price by area. This provides a flexible and economically coherent representation of parcel-level valuation while acknowledging the inherent complexity of land markets.

### INTERACTIONS

The final specification does not incorporate an extensive set of interaction terms between explanatory variables. Interaction terms allow the effect of one characteristic to vary depending on the level of another (for example, allowing the impact of parcel size to differ across locations or property types). Such terms can improve model performance by capturing non-linear and context-specific relationships.

In principle, the inclusion of carefully selected interaction terms could increase explanatory power and allow more detailed modelling of scale effects and variation across different types of parcels and locations. However, interaction structures substantially increase model complexity and parameter dimensionality, particularly in a large dataset with multiple categorical controls. This can reduce interpretability, increase the risk of overfitting, and weaken coefficient stability across subsamples.

The modelling approach therefore prioritised a parsimonious and stable specification that performs consistently across the dataset while remaining transparent and interpretable. The results should be viewed within this modelling choice.

### SUMMARY

Taken together, the diagnostic evidence indicates that the model provides stable and proportionate parcel-level land valuations across a wide range of transaction values and market contexts. While no empirical model can fully capture all site-specific and transactional factors influencing land prices, the results demonstrate that the chosen specification performs consistently, remains robust across subsamples, exhibits no material systematic pattern of prediction error, and remains appropriately transparent about modelling limitations. The model therefore provides a robust analytical

foundation for parcel-level land valuation within the scope and limitations outlined above.[4]

## LAND VALUATION USING THE HEDONIC MODEL

In order to estimate land values, the parameters on structure attributes were set equal to zero, as per Equation 2. Summary statistics for land valuation for parcels in the Transactions Database are shown in Table 3.1.3.

**Table 3.1.3: Lot 1 land value estimates for the Transactions Database**

| Minimum | Median | Mean | Max |
|---|---|---|---|
| £67,937 | £204,411 | £219,202 | £784,407 |

We also used the model to generate out-of-sample land-value estimates for every parcel in both the National Land Parcel Database and the LSOA Land Parcel Database.

Figure 3.1.1 illustrates the average land value per parcel in each LSOA, based on parcels in the National Land Parcel Database. Clicking on the image will take you to an interactive, scrollable version of this map.

As can be seen, the lowest land values per parcel are found predominantly in the south Wales valleys but with pockets of low values elsewhere, including in Swansea, Newport, Wrexham, Ffestiniog, Holyhead, Rhyl and Pembroke Dock. Most of the highest land values per parcel are found in an arc in the southeast that runs through the Vale, Cardiff, Monmouthshire and up into Brecknock. There are also pockets of high value around the Mumbles. The stark differences between nearby LSOAs is notable. For example, central Colwyn Bay (Glyn 2) is amongst the lowest 10% of average land values per parcel in this model, whilst neighbouring Colwyn Heights (Rhiw 1) is in the highest 10%. These patterns are consistent with those reported in ap Gwilym et al (2020).

Table 3.1.4 summarises the land values estimated for parcels in the LSOA Land Parcel Database. These summary statistics should be interpreted with caution given the significant heterogeneity in the nature of land parcels across the nine LSOAs within this database.

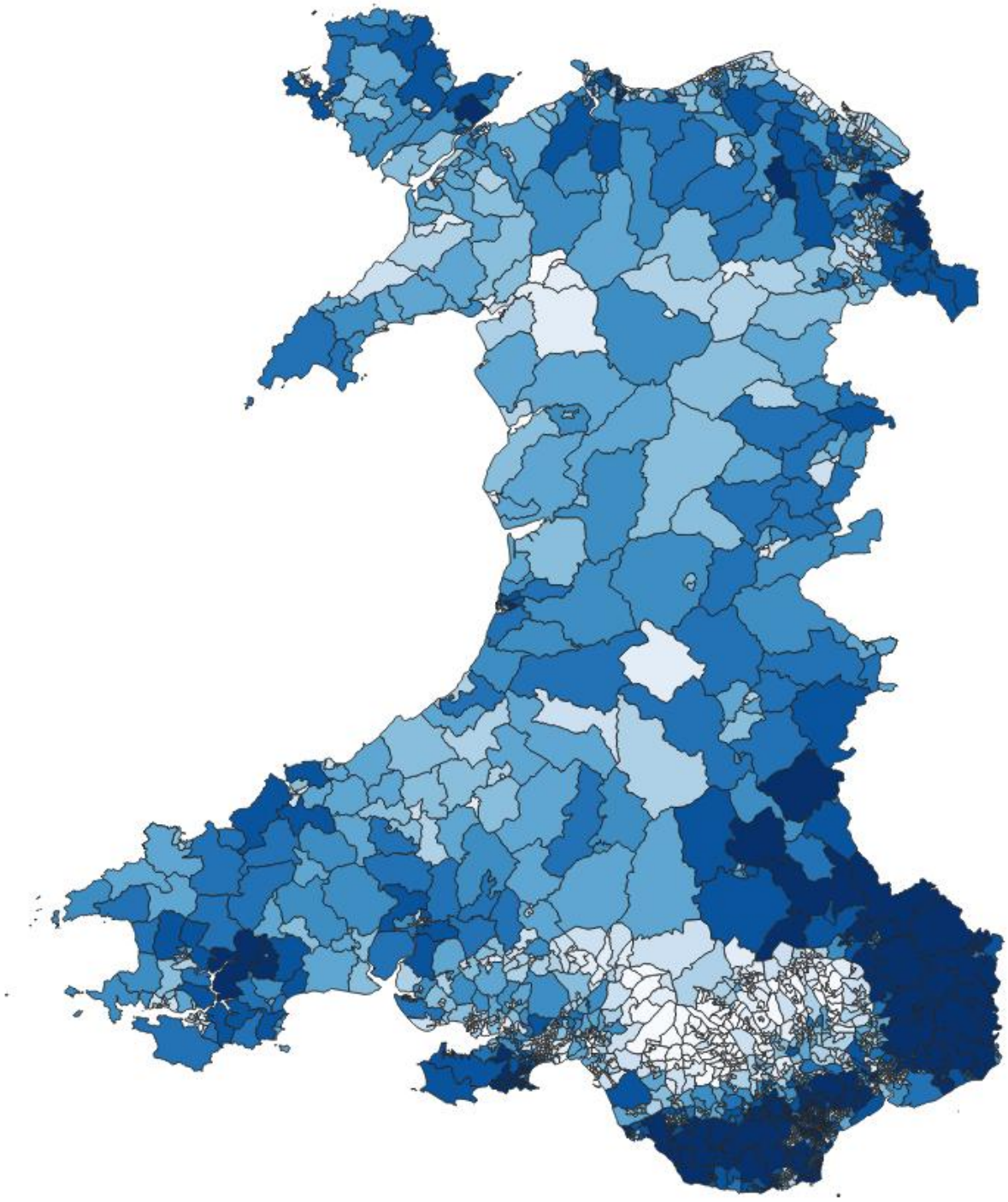
[4] Further details can be found in the Lot 1 technical assessment in Appendix B

**Table 3.1.4: Lot 1 land value estimates for the LSOA Land Parcel Database**

| LSOA | Number of observations | Median land value | Mean land value |
|---|---|---|---|
| **Trawsfynydd** | 5,652 | £167,258 | £245,375 |
| **Broughton North East** | 1,639 | £228,359 | £285,466 |
| **Knighton 2** | 1,280 | £191,413 | £242,939 |
| **Aberystwyth Rheidol 1** | 616 | £212,765 | £225,875 |
| **Scleddau** | 4,105 | £213,820 | £299,652 |
| **Porthcawl East Central 2** | 1,038 | £200,702 | £218,334 |
| **Rhigos** | 3,961 | £184,303 | £264,776 |
| **Usk 2** | 1,070 | £395,150 | £468,369 |
| **Cathays 12** | 1,119 | £307,152 | £378,116 |

Figure 3.1.2 presents thumbnail images of the results of Lot 1 for the nine identified LSOAs. It is important to note that the nine thumbnails are presented at different scales in Figure 3.1.2, which can give a false initial impression. For example, individual terraced properties can be discerned in the maps of Aberystwyth Rheidol 1 and Porthcawl East Central 2 but this is simply not possible in the case of Trawsfynydd or Scleddau. We recommend that you click on the images, which will take you to an interactive, scrollable map showing all nine LSOAs at a consistent scale.

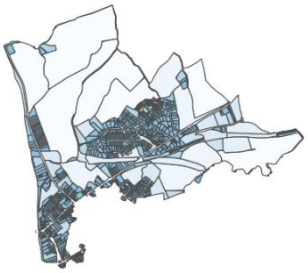
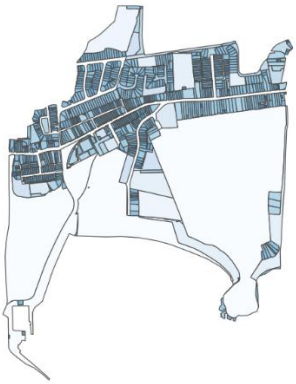
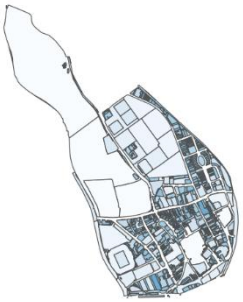
**Figure 3.1.1: Lot 1 average land values per parcel in the National Land Parcel Database by LSOA**



Darker shades of blue represent higher land values per parcel.

Please click on the map to access an interactive, scrollable version.

**Figure 3.1.2: Lot 1 land values per m$^2$ at the parcel level**

| | | |
|---|---|---|
| Trawsfynydd:  | Broughton North East:  | Knighton 2:  |
| Aberystwyth Rheidol 1:  | Rhigos:  | Usk 2:  |
| Scleddau:  | Porthcawl East Central 2:  | Cathays 12:  |

Darker shades of blue represent higher land values per parcel.

Please click on the map to access an interactive, scrollable version.

## LOT 2: ADVANCED ALGORITHMIC AND MACHINE-LEARNING APPLICATIONS

### INITIAL MODELLING

Initial modelling results yielded a large Root Mean Squared Error, with units in pounds, which is easier to conceptualise. Note:

- The cross-validated error is always larger as it represents a more accurate view of future performance.
- The additional interactions identified from decision tree analysis have notably improved the model (2nd row drop in error)
- Simply adding the interactions with all the variables might seem like it improves the model with a lower RMSE, but it is actually an overfit model that is likely to produce worse estimates in the future with a higher cross-validated RMSE. This highlights the value of the Lasso regression.

**Table 3.2.1: Results of the baseline LASSO model**

| Model | RMSE | 10-fold Cross-Validated RMSE |
|---|---|---|
| LASSO (All Variables) | £229,365 | £231,387 |
| LASSO (7 identified Interactions) | £210,192 | £217,921 |
| Linear Regression with LASSO/Tree Interactions | £205,856 | £353,523 |

### IMPROVED MODELLING

Statistical analysis of the pricing data and model residuals revealed that large outliers were driving the results presented above. These not only inflate the expected error, but also make the model substantially worse by focusing too much on rare, unusual cases; furthermore, these cases are even rarer today and in our key LSOAs. With the further assistance of decision tree analysis, outliers were removed, specifically those in the top 0.1% of the highest prices and those with a Floor Area greater than 8418 m². This totalled 1,018 properties removed, which, a priori, we thought would improve the modelling, and the results below confirm with RMSEs below £100,000.

**Table 3.2.2: Results of the LASSO model with outliers removed**

| Model | RMSE | 10-fold Cross-Validated RMSE |
|---|---|---|
| LASSO (All Variables) | £98,536 | £98,610 |
| LASSO (with 8 identified Interactions) | £93,015 | £93,480 |
| Linear Regression with LASSO/Tree Interactions | £92,989 | £98,324 |

The relative findings from the initial modelling hold: the interactions between variables notably improved the Lasso model. 8 interactions identified via decision tree analysis and confirmed by Lasso regression are shown below. These can be viewed as having a combined effect on price. For example, the effect of Floor Area on Price varies by Property Type, which is a very logical result. We clearly see that a few variables (Floor Area, Property Type, Property Type Category, Main Structure Area, Freehold Parcel Area, and Local Authority) are consistently represented in these significant interaction terms.

**Table 3.2.3: Identified interactions in the LASSO model with outliers removed**

| Variable 1 | Variable 2 |
|---|---|
| Floor Area | Property Type (Flats/ Terraced etc.) |
| Property Type Category (Retail, Utilities, Healthcare, Hospiltality etc.) | Freehold Parcel Area |
| Main Structure Area | Freehold Parcel Area |
| Floor Area | Local Authority |
| Floor Area | Property Type Category (Retail, Utilities, Healthcare, Hospiltality etc.) |
| Floor Level | Local Authority |
| Main Structure Area | Property Type (Flats/ Terraced etc.) |
| Floor Area | Freehold Parcel Area |

## LOG TRANSFORMATIONS

Although removing outliers greatly improved model performance, further residual analysis revealed that our raw linear and LASSO models consistently undervalue some price ranges while overvaluing others. When dealing with large numerical variables (positive skewness), a common approach is to model the log of the variable; that is, predict log(price) rather than price. This transformation preserves the ordering of house prices while compressing them into a more manageable form, and it can easily be undone to calculate final price predictions. This led to better modelling of residential properties, reducing the overall cross-validated RMSE from **£93,480** to **£88,063**. After a thorough investigation, all large numerical variables related to area (floor area, main and other structure area, freehold parcel area) were also transformed during this step. This updated modelling could be undervaluing large parcels of land due to the log transformation, but it is preferred because it produces much better predictions in most instances.

**Table 3.2.4: Results of the LASSO model with log transformations**

| Model | RMSE | 10-fold Cross-Validated RMSE |
|---|---|---|
| LASSO (All Variables) | £91,137 | £91,172 |
| LASSO (with 8 identified Interactions) | £88,021 | £88,063 |
| Linear Regression with LASSO/Tree Interactions | £87,983 | £88,145 |

## FURTHER IMPROVED MODELLING – THE FINAL CHOSEN MODEL

Based on our knowledge of the data and the above results, which feature Property Type multiple times, we decided to model residential properties separately from non-residential properties. This alone reduced the RMSE to £87k, but also presented an opportunity to refine the interaction terms separately for residential and non-residential. To assist in identifying potential interactions, machine-learning-based Bayesian Networks were used. Unique networks were separately trained on residential and non-residential datasets, and candidate interactions were identified from the relationships within each network. Specifically, if two variables were related both to each other and to price, then this interaction was trialled in the final model. This further improved the combined results; Lasso with interactions was once again the best model:

- The Lasso with interactions model 10-fold cross-validated RMSE dropped from £88,063 to **£85,082**.

The model for residential properties (1,229,749 out of 1,243,661 data points) was more accurate with an RMSE of **£82,754**, revealing that we are notably better at predicting those properties. The key interactions for residential properties are listed in Table 3.2.5 below.

The model for non-residential properties (only 13,912 out of 1,243,661 data points) was far less accurate, with an RMSE of **£204,737**, which indicates they are harder to predict. This is not surprising given the relatively small number of them. In further contrast with residential properties, log transformations did not improve this model. Thus, log transformations were only used for the residential model. The key interactions for non-residential properties are listed below. Interestingly, duration (a variable from the PPD reflecting tenure) occurs in almost all significant interactions for non-residential properties, but not for residential properties.

**Table 3.2.5: Identified interactions in the model for residential properties**

| Variable 1 | Variable 2 | Variable 3 |
|---|---|---|
| Floor Area | Property Type (Flats/ Terraced etc.) | % Adults with No Qualifications in LSOA |
| Main Structure Area | Property Type (Flats/ Terraced etc.) | Freehold Parcel Area |
| Floor Level | Floor Area | - |
| Number of Heated Rooms | Property Type (Flats/ Terraced etc.) | - |
| Floor Area | Property Type (Residential vs. Unknown) | - |
| Floor Area | Local Authority | - |
| Floor Area | Freehold Parcel Area | - |
| Local Authority | Distance to Train Station | - |
| Local Authority | Distance to Arts Centre | - |
| Local Authority | Key Stage 4 Average Score (LSOA) | - |
| Local Authority | Distance to Ice Rink | - |
| Local Authority | Long-term Illness Rate (LSOA) | - |
| Long-term Illness Rate (LSOA) | Key Stage 4 Average Score (LSOA) | - |

**Table 3.2.6: Identified interactions in the model for non-residential properties**

| Variable 1 | Variable 2 |
|---|---|
| **Duration (Freehold vs Leasehold)** | Local Authority |
| **Duration (Freehold vs Leasehold)** | Freehold Parcel Area |
| **Duration (Freehold vs Leasehold)** | Adults with No Qualifications in LSOA (%) |
| **Duration (Freehold vs Leasehold)** | Floor Area |
| **Duration (Freehold vs Leasehold)** | Crime (Theft) Rate |
| **Duration (Freehold vs Leasehold)** | Distance to In-Town Retail |
| **Duration (Freehold vs Leasehold)** | Property Type (Flats vs. Terraced etc.) |
| **Floor Area** | Freehold Parcel Area |

## OTHER MODELS – SVR AND KNN

For robustness checks, alternate machine learning models identified in our proposal from existing literature were created. Support Vector Regression (SVR) with a linear kernel resulted in a cross-validated RMSE of **£102,688**, while k-Nearest Neighbours (kNN) with k=7 resulted in a RMSE of **£100,043**. These results are substantially short of our best LASSO models (**£85,082**), and so no change is made to our final model.

kNN models are also known to be good at predicting data points that would be outliers for other models. For robustness, the highest 0.1% of residuals across all property types were predicted by a kNN model trained on the rest of the dataset. This kNN model outperformed the LASSO models on 1003/1242 (80%) of the tested properties. This shows that if we had a large dataset of land-only sales, kNN would likely be the best option for valuation. However, this is not plausible, and due to the nature of kNN, it cannot be easily used to predict land values from property prices as other models can.

## OTHER MODELS - ADVANCED NEURAL NETWORK AND XGBOOST MODELLING

We did not recommend such complex modelling because it is less stable for producing land-only estimates, as the structure and land components cannot be separated due to the model's complex structure (see land valuation section below for more details). However, we did perform the modelling to give an idea of how a more complex non-linear machine learning model would perform. Interestingly, the results are not an order of magnitude better than the above results, which gives us confidence that only minor improvements can be made with different modelling techniques; rather, improvements in the data would be required to achieve notable improvements in the modelling. Further, the variation of the generated land valuations from such complex modelling was substantially higher, which matched our concern from the proposal and why we didn’t recommend them as the primary model. The specifics of this modelling follow.

We tested two advanced non-linear approaches to assess whether they can improve prediction accuracy. For the neural network (NN), removing only the most extreme high-price transactions (the top 0.1%, i.e., excluding values above the 99.9th percentile of the scaled price) and incorporating the key interaction effects previously identified via the decision tree analysis resulted in an RMSE of £72,661. The XGBoost model performed best overall, achieving an RMSE of £68,831, indicating substantially stronger predictive accuracy than the NN in this setting. XGBoost also outperformed the LASSO model**.** The table below reports the cross-validated RMSE per LSOA for the NN and XGBoost models, which provides the most reliable indication of expected performance on new data among the advanced models. This finding is consistent with the literature (Ma et al., 2020; Jafary et al., 2024).

**Table 3.2.7: RMSE by LSOA in NN and XGBoost models**

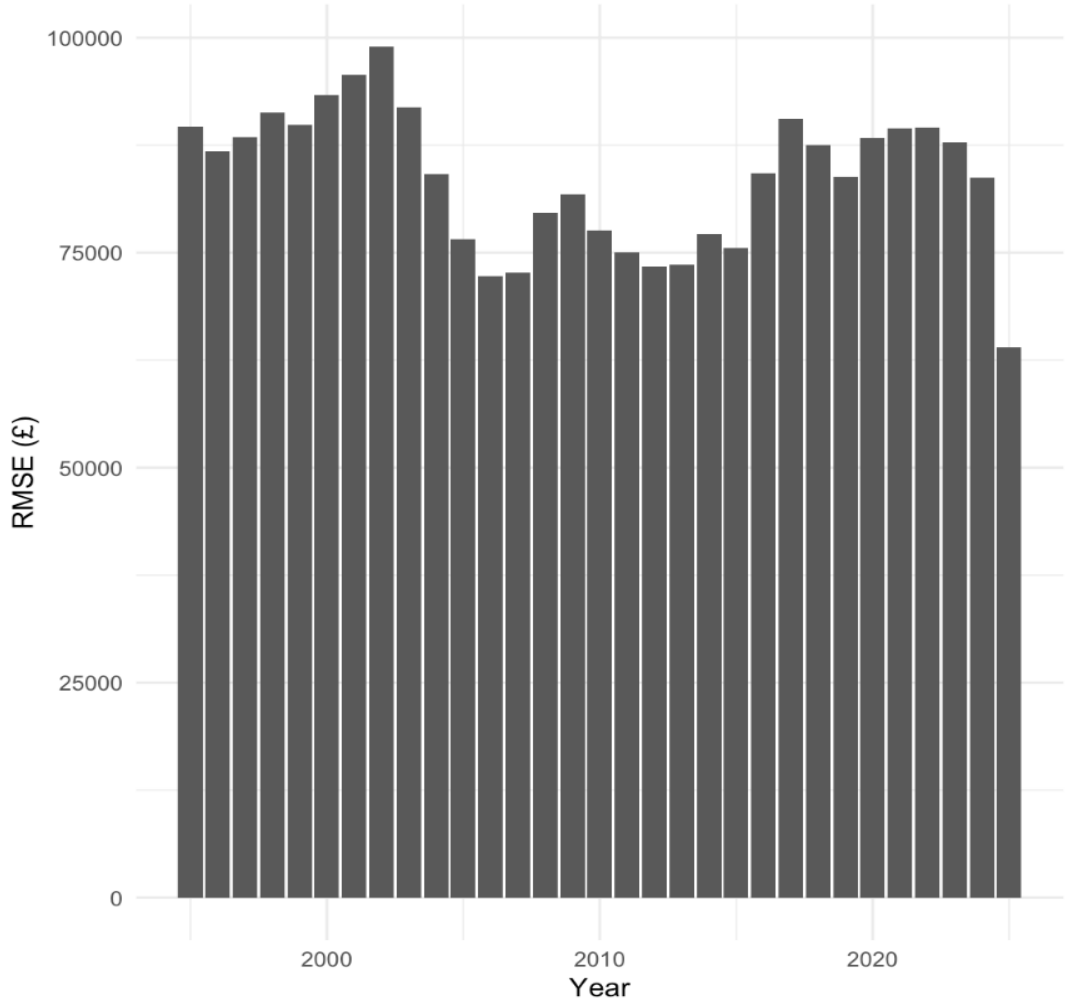
| LSOA ID | LSOA Name | NN RMSE | XGBoost RMSE |
|---|---|---|---|
| **W01000114** | Trawsfynydd | £64,541 | £41,466 |
| **W01000255** | Broughton North East | £94,742 | £84,281 |
| **W01000449** | Knighton 2 | £74,727 | £68,355 |
| **W01000517** | Aberystwyth Rheidol 1 | £73,735 | £72,552 |
| **W01000617** | Scleddau | £94,372 | £95,545 |
| **W01001045** | Porthcawl East Central 2 | £61,591 | £59,764 |
| **W01001233** | Rhigos | £98,667 | £104,435 |
| **W01001597** | Usk 2 | £115,061 | £109,063 |
| **W01002019** | Cathays 12 | £101,514 | £91,043 |

## FURTHER ANALYSIS OF THE FINAL LASSO MODEL

The final Lasso model, with interactions, outliers removed, and residential properties modelled separately with a log transformation, reveals no worrying trend over time. Figure 3.2.1 shows that the model improves in more recent years, which is encouraging.

Table 3.2.8 provides a breakdown of the model performance across the nine specified 9 LSOAs, revealing substantial variation, as to be expected given the large differences between LSOAs.

**Figure 3.2.1: Cross-validated RMSE by year of transaction**




Although we cannot quantify it precisely and so cannot be certain, we do believe that more of the error is coming from the structure, not the land component, so we believe the land component to be better modelled than these values. However, even if we factor that in, the errors remain concerningly high, indicating that this modelling should not be implemented in its current form. We believe that improved data quality and substantially more land-only sales would be needed to achieve a significant performance improvement that would make this modelling implementable.

**Table 3.2.8: RMSE by LSOA in the final model**

| LSOA ID | LSOA Name | Short Description | Cross-Validated RMSE |
|---|---|---|---|
| **W01000114** | Trawsfynydd | Rural villages in upland terrain. | £66,866 |
| **W01000255** | Broughton North East | Mix of residential, industrial and farmland. | £87,074 |
| **W01000449** | Knighton 2 | Small town and rural hinterland. | £77,025 |
| **W01000517** | Aberystwyth Rheidol 1 | Town centre: residential and retail. | £89,964 |
| **W01000617** | Scleddau | Mostly farmland with small, scattered settlements. | £108,451 |
| **W01001045** | Porthcawl East Central 2 | Medium-sized town. Residential and retail. | £66,507 |
| **W01001233** | Rhigos | Post-industrial, rural area. Industrial estate and villages | £103,475 |
| **W01001597** | Usk 2 | Small town, residential with some local services and green space. | £129,131 |
| **W01002019** | Cathays 12 | Cardiff city centre – includes stadium, retail areas, civic centre and parks. | £143,135 |

## LAND VALUATION USING OUR FINAL LASSO MODEL

As per Equation 2, to extract the land component from property prices, coefficients in the LASSO models related to structure were set to 0. This means that all structure variables would have no impact on land valuation, and the remaining component would be the value of the land itself. Summary statistics for land valuation on the training dataset are shown below, in Table 3.2.9. The effect of the log transformation on residential property prices is apparent here, as less extreme values are observed at both ends of the range. Although this is a limitation of the method, increased accuracy for most properties was a worthwhile trade-off.[5]

[5] Note that an equivalent decomposition using XGBoost feature contributions is less reliable, as baseline and interaction effects are not separable in the same way, and the residual after removing structural contributions can produce implausible negative land values (e.g. minimum land value of £-104,843 in the case of residential land and £-3,728,141 in the case of non-residential). Nevertheless, once a sufficiently large dataset of land attributes paired with reliable land-value observations (or validated land-value estimates) becomes available, XGBoost remains a promising and robust modelling option for capturing non-linear effects and complex spatial heterogeneity in land valuation.

**Table 3.2.9: Lot 2 and value estimates for the Transactions Database**

| Property Category | Minimum | Median | Mean | Max |
|---|---|---|---|---|
| **Residential** | £82,355 | £173,946 | £178,264 | £371,054 |
| **Non-Residential** | £25,000 | £147,082 | £166,412 | £1,417,889 |

We used the final model to generate out-of-sample land-value estimates for every parcel in both the National Land Parcel Database and the LSOA Land Parcel Database.

Figure 3.2.2 illustrates the average land value per parcel in each LSOA, based on parcels in the National Land Parcel Database. Clicking on the image will take you to an interactive, scrollable version of this map.

As can be seen, the lowest land values per parcel are concentrated in the south Wales valleys. The highest land values per parcel are less concentrated. They occur in parts of Cardiff, Swansea and their rural hinterlands; in various parts of Monmouthshire; and in coastal pockets such as Tenby and Beaumaris. These patterns are similar to those reported in ap Gwilym et al (2020) and in Lot 1 of this project.

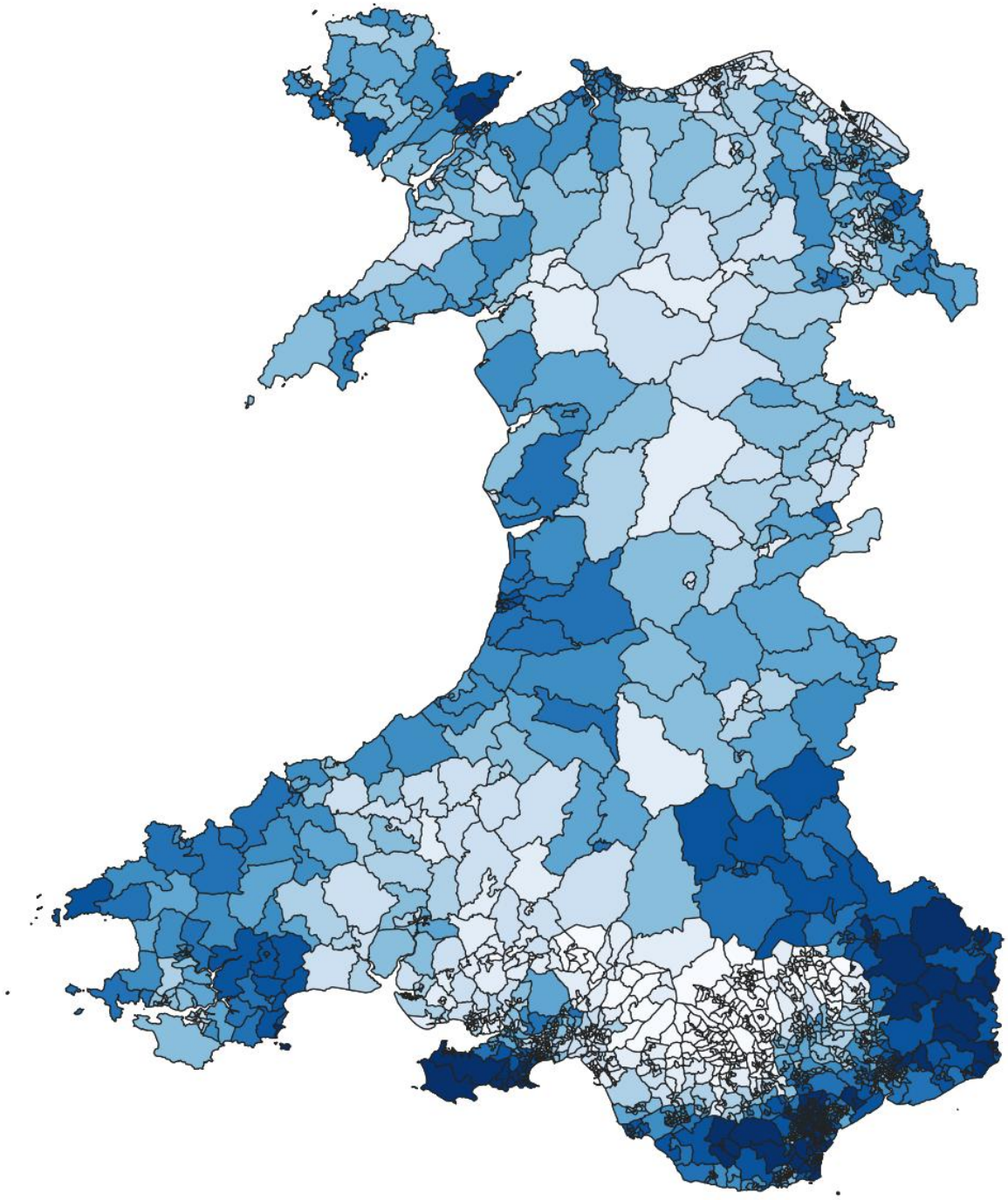
Table 3.2.10 summarises the land values estimated for parcels in the LSOA Land Parcel Database. These summary statistics should be interpreted with caution given the significant heterogeneity in the nature of land parcels across the nine LSOAs.

**Table 3.2.10 Lot 2 land value estimates for the LSOA Land Parcel Database**

| LSOA | Number of observations | Median land value | Mean land value |
|---|---|---|---|
| **Trawsfynydd** | 5,652 | £138,202 | £142,555 |
| **Broughton North East** | 1,639 | £177,109 | £189,936 |
| **Knighton 2** | 1,280 | £179,917 | £177,997 |
| **Aberystwyth Rheidol 1** | 616 | £228,700 | £227,969 |
| **Scleddau** | 4,105 | £194,432 | £197,901 |
| **Porthcawl East Central 2** | 1,038 | £202,016 | £199,970 |
| **Rhigos** | 3,961 | £107,993 | £113,681 |
| **Usk 2** | 1,070 | £264,429 | £266,122 |
| **Cathays 12** | 1,119 | £661,144 | £623,031 |

Figure 3.2.3 presents a much more detailed picture of the estimated land values. It shows thumbnail images of the results of Lot 2 for each land parcel the nine identified LSOAs.

**Figure 3.2.2: Lot 2 average land values per parcel in the National Land Parcel Database by LSOA**

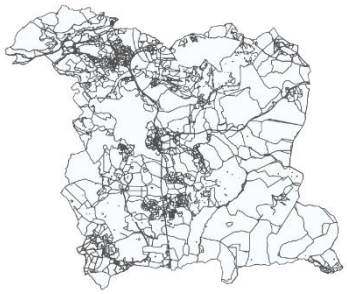
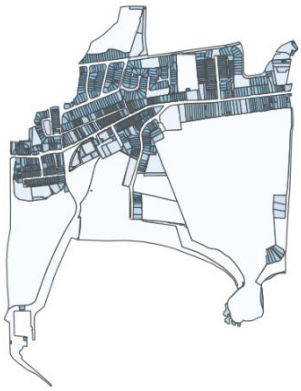


Darker shades of blue represent higher land values per parcel.

Please click on the map to access an interactive, scrollable version.

It is important to note that the nine thumbnails are presented at different scales in Figure 3.2.3, which can give a false initial impression. For example, individual terraced properties can be discerned in the maps of Aberystwyth Rheidol 1 and Porthcawl East Central 2 but this is simply not possible in the case of Trawsfynydd or Scleddau. We recommend that you click on the images, which will take you to an interactive, scrollable map showing all nine LSOAs at a consistent scale.

**Figure 3.2.3: Lot 2 land values per $m^2$ at the parcel level**

| | | |
|---|---|---|
| Trawsfynydd:  | Broughton North East:  | Knighton 2:  |
| Aberystwyth Rheidol 1:  | Rhigos:  | Usk 2:  |
| Scleddau:  | Porthcawl East Central 2:  | Cathays 12:  |

Darker shades of blue represent higher land values per parcel.

Please click on the map to access an interactive, scrollable version.

# LOT 3: FORMULA BASED VALUATION BY LAND AREA

This section presents the results from both strands of Lot 3. Strand A provides a comparative assessment of international formula-based land valuation systems. Strand B then uses Welsh transaction data to test a simplified, formula-based valuation approach consistent with the middle ground identified in Strand A—namely, the use of standardised location values rather than parcel-level spatial variables.

## STRAND A FINDINGS – INTERNATIONAL PRACTICE

This sub-section examines land-valuation practices in thirteen countries and sub-national systems, drawing out the key variables used, the rationale behind different methodological choices, and the implications for administrative capacity and policy effectiveness. The case studies are used to identify common themes and trade-offs, which then inform a set of lessons for Wales as it considers the potential development or reform of its own land-valuation arrangements.

## CASE STUDIES

### GERMANY – FEDERAL "BUNDESMODELL"

The German federal *Bundesmodell* provides a market-aligned but structurally simple approach to valuing land for public purposes. Under this model, the land component of value is determined exclusively by multiplying parcel area by an officially published standard land value (Bodenrichtwert). This applies both to *undeveloped land* and to the *land component of developed property*. The rules are statutory and uniform across most federal states that have not opted out into their own models (see case studies below). In the case of undeveloped land, the formula is an unambiguous:

**Land value = parcel area (m²) × standard land value (€/m²)**

For developed plots, the same basic structure applies but can include a small number of plot-related adjustment factors, such as plot depth or development intensity, referred to as the *Umrechnungsfaktor (U)*. This preserves a direct, transparent link between the physical land parcel and its assessed value, while embedding a degree of uniform market sensitivity through the use of standard land values compiled by local expert committees (*Gutachterausschüsse*).

Further information can be found on the German Federal Ministry of Finance website (Bundesministerium der Finanzen, no date).

**Table 3.3.1: Land valuation in the German federal "Bundesmodell"**

| Variable | Description |
|---|---|
| Size / area–related variables | |
| Parcel area (Grundstücksfläche, m²) | Total land area of the plot recorded in the cadastre. |
| Location and use–related variables | |
| Standard land value per square metre (Bodenrichtwert, €/m²) | Official standard land value for the location. |
| Plot adjustment factor (Umrechnungsfaktor U) | Adjustment factor for plot characteristics (e.g. depth, floor-space index) |
| Categorical selectors (embedded in Bodenrichtwert) | |
| Land-use / zoning category | Category such as residential, commercial or industrial, and the zoning of the parcel. It determines which Bodenrichtwert table / zone applies, but does not appear as a separate numeric regressor. |

### BAVARIA, GERMANY - "FLÄCHENMODELL"

The Bavarian Flächenmodell represents one of the simplest statutory approaches to land valuation found across the case studies. The model assigns a fixed monetary weight per square metre of land, without any reference to market values or standard land values (Bodenrichtwerte). The land component is therefore determined by a single, transparent calculation:

**Land value = land area (m²) × €0.04 per m².**

This structure ensures that the valuation depends solely on cadastral area and is entirely independent of local price dynamics. Unlike the German federal model, no adjustment factors or locational coefficients apply.

Further information on the statutory basis of the Bavarian Flächenmodell can be found on the official Bavarian Government legal portal (Bayerische Staatsregierung, no date).

**Table 3.3.2: Land valuation in the Bavarian "Flächenmodell"**

| Variable | Description |
|---|---|
| property-specific variables | |
| Land area (Grundstücksfläche, m²) | Total land area of the plot recorded in the cadastre. |

### HAMBURG, HESSE AND LOWER SAXONY, GERMANY - "WOHNLAGENMODELL"

The Hamburg, Hesse and Lower Saxony Wohnlagenmodell adapts the same underlying logic as Bavaria's area-based approach – assigning a fixed monetary weight of €0.04 per m² of land – but introduces location adjustments to reflect differences in land desirability while retaining structural simplicity. Hamburg applies the location effect through categorical residential location classes (simple, normal, good), while Hesse and Lower Saxony apply a continuous location factor based on the ratio of the parcel's standard land value (BRW) to the municipal average (dBRW), raised to the power of 0.3. This produces a basic land-valuation structure of:

**Land value = land area (m²) × €0.04 × location adjustment,**

with the variables (land area, BRW, dBRW, location category) summarised in Table 3.3.3.

Further details of each state's implementation can be verified through official government sources: Finanzbehörde Hamburg (no date), Hessisches Ministerium der Finanzen (no date) and Landesamt für Steuern Niedersachsen (no date).

**Table 3.3.3: Land valuation in the "Wohnlagenmodell" states**

| Variable | Description |
|---|---|
| Property-specific size variables | |
| Land area (m²) | Total land area of the plot recorded in the cadastre. |
| Standard land value (BRW) | Standard land value for the parcel (€/m²), used in the location factor in Hesse and Lower Saxony. |
| Average municipal standard land value (dBRW) | Average standard land value in the municipality (€/m²), used as the benchmark in the location factor. |
| Residential location category (Hamburg) | Categorical classification of residential areas (e.g. simple / normal / good), affecting the effective assessment rate for residential property. |

### DENMARK

The Danish system represents a value-based approach in which land is assessed according to its estimated market value in an undeveloped state (grundværdi). Land valuation forms part of the public property assessment, where variables such as parcel area, location (municipality, neighbourhood, address), and land-use/zoning category determine the land component of value. These variables are summarised in Table 3.3.4. In contrast to area-based models, Denmark does not embed land area directly in the statutory formula; instead, area and location feed into a mass appraisal

model that generates the assessed land value, which then serves as the base for land-related public policy uses.

Further information is available from the official Danish Property Assessment Agency (Vurderingsstyrelsen), which explains that land tax and property tax in Denmark are based on public property assessments, combining land and building valuations within a unified system. The agency notes that property owners pay land tax on the value of their undeveloped land, and that these values are generated through the national assessment framework, which incorporates market indicators and parcel-specific characteristics (Danish Property Assessment Agency, no date).

**Table 3.3.4: Land valuation in Denmark**

| Variable | Description |
|---|---|
| Land-valuation variables (assessment layer) | |
| Parcel / land area (m²) | Area of the plot; key explanatory variable in the land valuation model. |
| Location | Municipality, neighbourhood and address, used to determine local unit values and capture accessibility and amenity. |
| Land-use / zoning category | Permitted use (e.g. residential, commercial, agricultural), influencing the unit land value. |
| Tax-formula variables (land tax layer) | |
| Assessed land value (grundværdi) | Publicly assessed value of the land in undeveloped state, serving as the base for land tax. |
| Fraction of land value | Proportion of assessed land value used as the land tax base (commonly around 80%) |
| Municipal land tax rate | Municipal land tax rate applied to the land tax base |

## POLAND

Poland operates one of the clearest examples of a pure area-based land valuation system among the case studies. The statutory basis for land valuation is exceptionally straightforward: the taxable land area (in m²) is multiplied by a municipal rate per square metre, with no reference to market values, standard land values, or location adjustments. For the purposes of valuing land within public policy, this means that the sole quantitative driver of the land component is land area, while categorical distinctions – such as land used for business versus residential land – simply determine which rate applies. The variables captured in Table 3.3.5 (land area, land-use category, object type, and the municipal rate) define the parameters of the system.

This system reflects an underlying policy preference for maximum simplicity, predictability and administrative ease, rather than sensitivity to local land markets.

Since the formula does not react to price dynamics or spatial amenity differences, it is easy to administer and easy for landowners and municipal officials to understand. At the same time, this simplicity limits the usefulness of Polish land valuations for broader public-policy applications that require spatially nuanced information (e.g. land-value capture, planning viability analysis, or compensation). Nonetheless, the Polish model demonstrates the durability and interpretability of a strict “area × rate” valuation approach, which stands in marked contrast to the market-aligned or hybrid unit-value systems seen elsewhere.

The statutory basis for this area-based approach is set out in the Act of 12 January 1991 on Local Taxes and Fees (u.p.o.l.), which defines land tax as being assessed solely on land area, and is summarised for the public on the official Biznes.gov.pl portal.

**Table 3.3.5: Land valuation in Poland**

| Variable | Description |
|---|---|
| Property-specific physical variables | |
| Land area (A_land, m²) | Surface area of the taxable land parcel, in square metres |
| Classification and use variables | |
| Type of object | Land vs building vs structure |
| Use category | Category such as land used for business, residential land, or other land, determining the applicable rate per m². |
| Jurisdiction-specific rate variables | |
| Municipal rate per m² for land (r_land) | Annual rate per m² set by the municipality within national maxima. |

### MONTANA, USA

Montana’s valuation of Class Three agricultural land is a clear example of a productivity-based use-value system, designed to insulate agricultural property from market speculation. The value per acre is determined by capitalising net agricultural income using a statutory capitalisation rate of 6.4%, consistent with the legislative intent in the Montana Code Annotated §15-7-201, which requires agricultural land to be valued according to productive capacity rather than market price. Net income per acre is derived from soil productivity indices, allowable costs and commodity prices, with parcel value obtained by multiplying these subclass-specific per-acre values by the number of acres in each subclass (irrigated, non-irrigated, hay land, grazing, etc.), reflecting §15-6-133 MCA on Class Three property categories.

Montana's statewide cadastral system, maintained through the Montana State Library, displays land value and buildings value separately. For non-agricultural land (residential, commercial, industrial), the Department of Revenue values property using market value as required under §15-8-111 MCA, applying the three standard mass-appraisal approaches: cost, income and sales comparison. The sales-comparison method—called the MKT (Market) method within the cadastral viewer—is based on verified arm's-length sales and the Department's statistical modelling of comparable properties, as described in the official appraisal-process documentation, which sets out the cost, sales comparison, and income approaches as the core valuation tools.

**Table 3.3.6: Land valuation in Montana (class three agricultural land)**

| | |
|---|---|
| Property-specific physical & use variables | |
| Parcel size (acres) | number of acres in each agricultural subclass |
| Land use subclass | Irrigated farmland, non-irrigated cropland, hay land, grazing land, non-qualified agricultural land |
| Soil productivity indices | bushels of spring wheat per acre for cropland, animal-unit months per acre for grazing land, drawn from NRCS soil surveys |
| Irrigation status / water regime | Determines whether irrigated formulas (including water costs) apply |
| Income and cost parameters (policy/market inputs) | |
| Average commodity prices | for the base period (by crop/livestock type) |
| Typical crop-share or livestock-share percentage | e.g. 25% owner's share |
| Allowable water costs per acre | for irrigated land |
| Other production cost assumptions | fuel, inputs, etc. |
| Valuation and tax parameters | |
| Capitalisation rate R | statutory 6.4% unless changed |
| Class three tax rate $\tau_{\text{class 3}}$ | 2.16% of productive capacity value |
| Local mill levy | sum of local tax rates applied to the taxable value |

## ESTONIA

Estonia operates a national land valuation system in which the taxable value of land is determined through regular mass valuations conducted under the Land Valuation Act. Each cadastral unit receives a value derived from a per-square-metre (or per-hectare) unit value assigned according to the parcel's intended use, location, and physical attributes (e.g., land-use type, land composition, restrictions), as shown in Table 3.3.7. The result is a taxable value = unit value × cadastral area, which

forms the basis for multiple public-policy applications including planning, compensation, land consolidation and taxation.

Official sources confirm that this system is grounded in statute. The Land Valuation Act specifies that land valuation determines the “usual value of land”, using internationally recognised methods including sales comparison, income and cost approaches, and that mass valuation is the default mechanism for assigning value across all land parcels. The Land Tax Act links these mass-appraised values directly to the land tax base by stating that land tax is levied on the taxable value of land as determined through the Land Valuation Act’s procedures. The cadastral system maintained by the Estonian Land Board records parcel boundaries, intended purpose, land-use type, restrictions and taxable value, providing the authoritative spatial backbone for these valuations.

These arrangements are set out in the Land Valuation Act and Land Tax Act (Riigikogu 2022a,b), with the Estonian Land Board’s (2025) cadastral system providing the authoritative spatial and valuation data used in mass valuation and taxation.

**Table 3.3.7: Land valuation in Estonia**

| | |
|---|---|
| Cadastral and physical variables | |
| Cadastral area | Area of the cadastral unit (m² or ha) |
| Intended purpose (*sihtotstarve*) | residential, commercial, agricultural, forest |
| Land use type category | yard land, arable land, pasture, forest |
| Location | |
| Municipality | Local government area in which the cadastral unit is located. |
| Settlement / neighbourhood | Settlement or urban district, used in the mass valuation to differentiate land values. |
| Distance to centre | Variables capturing proximity to centres and accessibility (e.g. to main roads or services). |
| Other cadastral attributes | |
| Land composition | Land cover / composition |
| Restrictions | Easements, protection zones and other legal restrictions affecting land value. |

## JAPAN

Japan’s Rosenka framework offers one of the most transparent unit-value approaches to land valuation among the case studies. The National Tax Agency (2025a, b) publishes an annual roadside land value for each road segment, representing a standard per-m² unit value typically set at around 80% of market

value. For parcels fronting these roads, land valuation is calculated mechanically as land area × roadside value, with additional plot-adjustment factors applied to reflect parcel depth, shape, and frontage conditions. These variables are summarised in Table 3.3.8.

**Table 3.3.8: Land valuation in Japan (Rosenka system)**

| | |
|---|---|
| Property-specific physical & use variables | |
| Land area (m²) | Registered parcel area of the plot |
| Frontage road segment | Identification of the road (road code/segment) in front of the parcel; determines which rosenka (roadside value per m²) applies |
| Use / zoning category | Main use or zoning of the land |
| Plot configuration | Basic information on whether the parcel is interior, corner lot, flag lot, etc. |
| Unit land value & adjustment parameters | |
| Roadside value (rosenka, ¥/m²) | Standard value per square metre for land fronting a given road segment, published annually by the National Tax Agency |
| Lot-depth / shape correction factor | Multiplicative factor adjusting the standard rosenka for lots that are deeper/shallower or irregularly shaped compared with the standard model lot |
| Corner-lot / multiple-frontage factor | Factor applied when a parcel fronts more than one road (corner lots, through lots), typically increasing the land value to reflect better frontage |

The resulting valuation method is highly consistent across Japan, allowing the land component of property value to be determined with minimal discretion and high public visibility. While this system is used primarily for inheritance and gift taxation, the clarity of the unit-value × area × adjustment-factor structure makes Rosenka a widely used reference point in broader public-policy contexts, including urban land economics and planning feasibility analysis. Because buildings are typically valued separately using their own appraisal rules, the Rosenka system provides a pure land valuation based on standardised, nationally published values, ensuring comparability across regions while embedding local nuance through the detailed spatial granularity of road-segment values and correction coefficients.

## AUSTRALIA

New South Wales and Victoria operate state-level land valuation systems in which land is assessed according to its unimproved or site value, derived through market-based mass appraisal undertaken by each state's Valuer-General (PWC 2019, Revenue NSW 2025, State Revenue Office Victoria 2025). The statutory tax

formulas themselves do not include land-area or location variables; instead, they rely entirely on the assessed land value, meaning that the role of parcel-level characteristics sits squarely within the valuation stage rather than the tax formula. The mass-valuation models used by the Valuer-General incorporate site/parcel area, location (municipality, neighbourhood, distance to CBD, amenity access), zoning/permissible use, and site characteristics such as slope, corner-plot status and parcel shape, as summarised in Table 3.3.9.

In practice, parcel area is a central explanatory variable because comparable sales are routinely analysed on a per-square-metre basis, and market evidence forms the foundation of unimproved/site value estimates. NSW bases land tax on the three-year average of unimproved land values, while Victoria applies its site value as determined through the annual general valuation process, both embedding a consistent and market-aligned valuation framework. Although buildings do not enter the taxable land value, their presence influences market sales, which in turn shape the mass-appraisal outputs used to determine site value. This places NSW and Victoria firmly in the group of jurisdictions—along with Denmark, the Netherlands and South Africa—where valuation relies on comprehensive market analysis, but the tax system remains structurally simple by using only the final assessed land value as the tax base.

**Table 3.3.9: Land valuation in Australia (NSW and Victoria)**

| Physical and locational variables used in valuation | |
|---|---|
| Parcel/site area (m²) | key determinant of land value, often via comparable sales per m² |
| Location | municipality, local government area, neighbourhood, distance to CBD and amenities |
| Zoning / permissible use | residential, commercial, industrial, etc |
| Topography and site characteristics | slope, corner plot, shape and other factors affecting market value. |

## NETHERLANDS

The Netherlands applies a value-based property valuation system centred on the statutory WOZ value (*waarde onroerende zaken*), which represents the estimated market value of each property as of 1 January of the previous year. Municipal assessors determine this value through mass appraisal models that draw heavily on market sales, with key property characteristics – including plot size, location (municipality, neighbourhood, micro-location), building floor area, year of construction, and quality/condition – forming core explanatory variables. These inputs, particularly plot size, are summarised in Table 3.3.10, and they define how the land component is represented in the overall valuation framework.

Although the WOZ value incorporates both land and buildings, the modelling process treats plot size (grondoppervlakte/perceelgrootte) as a primary land variable for ground-bound dwellings, ensuring that land area explicitly influences the assessed market value. The resulting WOZ value is then used as the base for the Dutch municipal property tax (OZB), but the valuation itself serves much broader public-policy purposes, including national taxation, housing-market monitoring, and local planning. The Dutch system therefore exemplifies a market-aligned valuation regime in which land enters indirectly but decisively through the mass-appraisal model's attention to parcel area and spatial context, while the statutory tax formula remains structurally simple as OZB = municipal rate × WOZ value. (See Netherlands Enterprise Agency (RVO) 2022)

**Table 3.3.10: Land valuation in the Netherlands**

| Variables in the statutory tax formula | |
|---|---|
| WOZ value | Statutory property value (estimated market value) on 1 January of the previous year. |
| Municipal OZB rate | Percentage of WOZ value, set by each municipality, often with separate owner / user rates |
| Core physical and locational variables in the WOZ valuation model | |
| Living / floor area (*woonoppervlakte* or *totale gebruiksoppervlakte*) | Woonoppervlakte or totale gebruiksoppervlakte – total usable residential floor area of the dwelling; primary object characteristic but building-focused. |
| Plot size (*perceelgrootte / grondoppervlakte*) | Size of the land parcel; for ground-bound properties, plot size is explicitly recorded for every WOZ object and taken into account in the valuation. |
| Location | Municipality, neighbourhood and local situation (e.g. proximity to amenities, local market conditions), reflected through the choice of comparable sales and location variables in the model. |
| Year of construction | Primary characteristic affecting value (older vs newer dwellings). |
| Quality, condition and facilities | Secondary characteristics (maintenance, modernisation, level of amenities) that adjust the market value. |

## SOUTH AFRICA

South Africa operates a market-value–based property rating system under the Municipal Property Rates Act (MPRA) (Republic of South Africa 2004), in which municipalities levy property rates on the market value of land and improvements. The valuation basis is the open-market value a willing buyer would pay under normal conditions, with municipalities required to maintain and regularly update a statutory

valuation roll containing these values. Land-related variables are essential determinants of the market value recorded in the roll, and are summarised in Table 3.3.11.

Although buildings contribute to the total market value, the valuation process makes extensive use of GIS-based mass appraisal models, which incorporate spatial variables such as proximity to amenities, transport networks and local market conditions. The system thereby blends a market-aligned valuation framework with a structurally simple rating formula:

**property rates = rate-in-the-rand × rateable value.**

This positions South Africa alongside jurisdictions such as the Netherlands and Denmark, where valuation complexity is concentrated in the mass-appraisal stage, while the statutory instrument itself remains straightforward.

**Table 3.3.11: Land valuation in South Africa**

| | |
|---|---|
| Parcel size / site area | Land area (m² or hectares) used in sales comparison or cost approaches |
| Building / floor area | Gross building area, number of storeys, etc., for residential and commercial properties |
| Land use and zoning category | Residential vs business vs industrial vs agricultural, and specific zoning rights, which influence value and the applicable rate category |
| Location | Municipality, suburb, neighbourhood, and micro-location (proximity to CBD, transport, amenities); encoded in GIS and valuation models |
| Quality, age and condition of improvements | Age of buildings, quality of construction, maintenance level, and presence of outbuildings or infrastructure |

## ROMANIA

Romania operates an area-based land valuation system in which the tax base for land is defined exclusively by land area rather than market or unit-value assessments. The Fiscal Code (National Agency for Fiscal Administration 2015) specifies that land tax is calculated using fixed amounts per square metre, differentiated according to the rank of locality (0–V), zone within locality (A–D) and land-use category (construction land versus various agricultural land classes). This produces a transparent and easily administered formula:

**land tax = land area (m²) × Lei-per-m² tariff,**

requiring only cadastral area and basic categorical information, as summarised in Table 3.3.12.

This approach reflects a policy preference for administrative simplicity, ensuring that land valuation for taxation does not depend on market conditions or complex appraisal processes. Although used primarily for taxation, the model provides a consistent, nationally uniform basis for area-related public-policy decisions. Romania's land-tax system contains no market-value component, no unit land values and no valuation modelling; the only drivers of liability are spatial extent and the statutory per-m² tariff.

**Table 3.3.12: Land valuation in Romania (land tax)**

| Land tax – explicit area variables | |
|---|---|
| Land area (A_land) | Number of square metres of the parcel used for tax purposes. |
| Rank of locality | Classification of cities/communes into ranks (0, I, II, III, IV, V), affecting the per-m² amount. |
| Zone within locality | Zones A–D (central to peripheral) used in the per-m² tables |
| Category of land use | Construction land vs agricultural land (arable, pasture, vineyard, orchard, forest, land with water, etc.), each with specific amounts per m² and correction coefficients |

## BULGARIA

Bulgaria operates one of the most explicitly formula-driven land-valuation systems in the international sample. The Local Taxes and Fees Act (LTFA) (Republic of Bulgaria 2006) requires the tax evaluation price for land to be determined through a statutory formula combining a base tax value per m² with land area and a series of correction coefficients. These coefficients capture key contextual attributes including location, infrastructure availability, development zone and development characteristics, producing a valuation structure of the form:

**tax evaluation price = base value per m² × area × coefficients.**

The relevant variables are summarised in Table 3.3.13.

This system embeds both spatial and regulatory characteristics directly into the valuation function, making Bulgaria one of the clearest examples of a statutory assessment model rather than a market-based or mass-appraisal approach. Once the tax evaluation price is established, municipalities apply their own real estate tax rate – typically between 0.1% and 4.5% – to produce the final liability, leaving the underlying land-value calculation unchanged. Bulgaria's approach stands out for its transparent, formulaic structure, in which land valuation is reduced to measurable physical area and codified adjustment factors.

**Table 3.3.13: Land valuation in Bulgaria**

| | |
|---|---|
| Explicit area and price-per-m² variables | |
| Land area (m²) | Area of the land parcel (including built area); enters the land-assessment formula |
| Base tax value per m² (BGN/m²) | Different schedules for residential vs non-residential buildings and for different land types; serves as the base unit price for tax calculations |
| Correction coefficients | |
| Location coefficient (Cl) | Reflects the locality / zone where the property is situated (central vs peripheral, resort vs ordinary settlement, etc.) |
| Infrastructure coefficient (Ci) | Adjusts for the quality/availability of infrastructure (roads, utilities) |
| Development-zone coefficient (Cz) | Captures the planning/functional zone of the land |
| Development coefficient (Cd) and other coefficients | May reflect building height, number of floors, individual characteristics, obsolescence / age of the building, etc |

## LUXEMBOURG

Luxembourg's communal property tax (impôt foncier, IFON) is formally a value-based system built around a long-standing statutory structure in which the tax liability is calculated as valeur unitaire × taux d'assiette × communal rate. The valeur unitaire is an administratively determined unit value derived historically from 1941 rental values, updated through legal rules rather than market revaluation. The resulting structure does not explicitly incorporate land area or unit land values within the statutory formula, although parcel characteristics are embedded implicitly within the valuation methods applied by the Administration des Contributions Directes (ACD). These formal inputs are summarised in Table 3.3.14.

A major ongoing reform seeks to modernise IFON by shifting towards an explicit land-value basis (Chambre des Députés 2023). The government's reform model centres on a "base land value" derived from a unit land value per are/m², determined according to the parcel's PAG zoning category, development potential and distance to centres, and then multiplied by parcel area. The official simulator requires taxpayers to input parcel size and PAG zoning, making land area and land-use designation the core determinants of the revised tax base.

**Table 3.3.14: Land valuation in Luxembourg**

| Current statutory formula variables | |
|---|---|
| Valeur unitaire | Administrative unit value of the property, historically derived from rental value (1941 base), updated via legal rules |
| Taux d'assiette | Assessment rate depending on property category; multiplied by valeur unitaire to obtain base d'assiette |
| Taux communal | Communal percentage applied to the base d'assiette to obtain the impôt foncier for each property. |
| Variables highlighted in the reform and simulator | |
| Parcel area (superficie, in ares) | Users must input the parcel area from the cadastre/geoportal; the simulator uses this as a direct factor in base land value. |
| Zoning / PAG zone | The zone du PAG (local land-use plan) for the parcel (e.g. MIX-u, residential, commercial, green zone) determines the unit base value and reflects development potential |
| Base land unit value | Implied variable in the reform: a per-are (or per-m²) base value assigned according to zoning, development potential, and distance to urban centres |
| Abatements / reductions | For owner-occupied or bequeathed properties, an abatement is subtracted from the base land value before applying the communal rate. |

## SLOVAK REPUBLIC

The Slovak Republic applies an explicitly area-based land valuation system under its local immovable property tax regime (Slovak Republic 2004). The land component of the tax base is determined by multiplying land area (m²) by a statutory or municipally adjusted value-per-m², differentiated by land-use category (e.g. arable land, vineyards, gardens, built-up areas). These inputs are summarised in Table 3.3.15, and this simple valuation structure is a defining feature of the Slovak system. The resulting land value is then multiplied by the local land-tax rate, often around 0.25% of the assessed land value, to produce the final liability.

This model embodies a strong commitment to administrative simplicity, with parcel area serving as the primary quantitative driver and land-use category providing the minimal differentiation required for fiscal and planning purposes. There is no market-value component in the statutory land-valuation formula, and no requirement for mass appraisal or unit-value modelling. Instead, municipalities adjust the value-per-m² or the local tax rate to calibrate burdens across land types and local conditions.

**Table 3.3.15: Land valuation in the Slovak Republic**

| Land component | |
|---|---|
| Land area (m²) | direct multiplicative factor in the tax base |
| Value per m² of land (v_land) | statutory or municipally adjusted amount by land category (arable land, vineyards, gardens, built-up areas, etc.) |
| Location / municipality | affects the applicable value per m² and/or the local rate; |
| Land-use category | agricultural vs other land types, with different values per m² |

## EMERGENT THEMES

Across the 13 jurisdictions examined, several clear themes emerge in how governments conceptualise, model and apply land valuation for public-policy purposes. Although the systems differ markedly in complexity and legal architecture, they converge around a set of recurrent design choices relating to the treatment of land area, the role of location, valuation methodology, and administrative feasibility. These themes provide a valuable comparative lens for reflecting on possible approaches within the Welsh context.

### DIFFERENT LEVELS OF MARKET INFLUENCE

This systematic review of international practice of 13 jurisdictions has identified three broad approaches to incorporating market prices into land valuation:

**1. Pure area-based models**

Used in Poland, Romania, Slovakia, and parts of Germany (Bavaria, Hamburg). These systems calculate land value (or tax base) as area × fixed amount, with optional categorical modifiers.

They are highly predictable, low-cost, and transparent, but unconnected to market values.

**2. Standardised land values**

Japan's Rosenka system, Luxembourg's proposed reform, Bulgaria's coefficient model, and Germany's federal Bundesmodell fall into this class.

They apply standardised unit land values, informed by market conditions, multiplied by area and small adjustment factors.

These systems improve market sensitivity while remaining more interpretable than full mass appraisal.

**3. Market-value systems using mass appraisal**

Denmark, the Netherlands, NSW/Victoria and South Africa rely on hedonic or comparable-sales modelling, with market value serving as the tax or assessment

base. These systems offer the closest alignment with real land markets but require strong administrative institutions, good data and routine revaluations.

### COMMON VARIABLES ACROSS SYSTEMS

A commonality across all jurisdictions is that land area is always a central variable, either explicitly embedded in the statutory formula or implicitly embedded in mass-valuation models, where area is a key predictor of market value.

Other common variables across the systems reviewed include:

- Unit land value (see next theme).
- Location categories. These include administrative areas, neighbourhoods, distances to amenities and other measures such as location class in Hamburg.
- Land-use or zoning categories.
- Measures of plot shape, development intensity, and infrastructure.

### DIVERGENT USES OF LOCATION

Location enters valuation systems in three broad ways.

#### 1. No or minimal location differentiation

Poland, Romania, Bavaria and other pure area-based regimes apply zero or near-zero location sensitivity, relying instead on municipality-wide or locality-wide flat tariffs or equivalence figures.

#### 2. Simple categorical adjustments

Some systems incorporate coarse location categories:

- Hamburg's "simple/normal/good" residential location classes;
- Romania's locality rank and intra-locality zones;
- Bulgaria's location, development-zone and infrastructure coefficients.

These retain simplicity while acknowledging spatial differentiation.

#### 3. Continuous, market-aligned location factors

Systems such as:

- Germany (Hesse and Lower Saxony), with BRW/dBRW^0.3 location factors,
- Denmark, where neighbourhood and address feed directly into valuation models,
- Netherlands and South Africa, where market sales implicitly encode spatial variation, demonstrate a more granular, continuously varying role for location in land valuation.

This continuum, from no location signal to fully market-driven location modelling, illustrates the trade-off between equity, administrative complexity and responsiveness to local market conditions.

## TRANSPARENCY VERSUS COMPLEXITY

The cases reveal a strong inverse relationship between valuation accuracy and interpretability:

- Simple systems (Bavaria, Romania, Poland) are easy to explain and predictable, but may generate inequities where land values diverge widely within or between localities.
- Market-based systems (Denmark, Netherlands, South Africa) better reflect spatial value differences but rely on complex models that can be opaque to taxpayers and policymakers.
- Hybrid systems (Japan, Germany's Bundesmodell, Luxembourg's reform model) strike a middle ground, offering explicit unit-value schedules that are both transparent and location-sensitive.

Transparency is especially important for public trust and for enabling policy uses beyond taxation, such as planning, compensation and value-capture analysis.

## CADASTRE QUALITY AND VALUATION INFRASTRUCTURE

Countries with robust mass-valuation systems (e.g., Denmark, Netherlands, NSW/Victoria, Estonia) all rely on:

- complete, accurate cadastral records,
- high-quality sales data,
- centralised valuation agencies, and
- periodic revaluations to maintain credibility.

By contrast, area-based countries typically have weaker valuation infrastructure or prioritise administrative simplicity over precision.

## DISTINGUISHING LAND FROM BUILDINGS

Across the case studies, land and buildings are often valued separately, but with different degrees of granularity:

- In Japan, land is valued via roadside unit prices, buildings via separate appraisal rules.
- In Germany's Bundesmodell, land and buildings have distinct formula components.
- In Montana, agricultural land values depend entirely on productive land capacity, so buildings are mostly irrelevant, but land and buildings are explicitly differentiated for residential and commercial properties.
- In value-based systems (Denmark, Netherlands), land–building splits emerge from the mass-appraisal model rather than statute.

This reinforces that land is conceptually separable from improvements, even where the tax or valuation system ultimately aggregates them.

## VARIED POLICY PURPOSES

Although taxation is the dominant driver of most valuation systems, several jurisdictions structure valuations to support additional public-policy goals:

- Land-use planning and zoning (Luxembourg, Netherlands, Estonia).
- Discouraging speculation or holding costs (Hamburg's location classes; Japan's use of Rosenka in inheritance tax; Romania's locality-zone structure).
- Agricultural land protection (Montana).

## LESSONS FOR WALES

The themes identified above offer a set of practical insights highly relevant to Wales as it considers the future of land-valuation arrangements for public-policy purposes. While each jurisdiction operates within its own legal, institutional and political context, several recurring patterns speak directly to the Welsh challenge of developing valuation approaches that are robust, transparent, equitable, and administratively feasible within a small-country governance structure.

### 1. Cadastral data is the foundation of any Welsh system

Across all international cases land area is always a core building block. For Wales, this confirms that:

- Cadastral accuracy, including clear parcel boundaries and area measurements, is a key prerequisite.
- Any future Welsh valuation system, whether area-based or market-oriented, will depend structurally on reliable area data.
- Siloed approaches to data, as currently exists, is a key barrier to implementing any valuation model Wales might choose.

This suggests that investment in geospatial infrastructure is foundational to later policy choices.

### 2. Wales must decide how much spatial differentiation it wants to administratively sustain

A system with no spatial differentiation would be simple but misaligned with the country's substantial variation in land values. A fully continuous, market-modelled approach would provide the most accuracy but also require substantial institutional capacity, frequent revaluations, high-quality data and extensive public-facing transparency mechanisms.

This is a strategic choice that must be made early, since location design largely determines administrative burden and political acceptability.

### 3. Wales must consider which broad approach to incorporating market prices into land valuation best fits its policy goals

A pure area-based model could be viable for specific public-policy uses (e.g., planning charges or certain forms of land-use regulation), but is unlikely to satisfy equity or market-awareness concerns for broader valuation roles. Standardised land values could be developed by implementing official land-value zones, potentially through a national Land Value Map maintained by a central body. Market-value mass-appraisal would require investment, multi-year development, and integrated datasets across planning, land registry, and local government.

### 4. Transparency needs to be balanced with precision

The international comparison shows that simplicity and interpretability often stand in tension with market accuracy.

### 5. Institutional development is key

Lessons from Estonia, the Netherlands and Denmark highlight the importance of high-quality cadastre and valuation infrastructure. Wales currently has:

- Fragmented local-authority data arrangements,
- Centralised but incomplete Land Registry datasets,
- Valuable Ordnance Survey data but with licensing and integration challenges,
- No standing national valuation office for land.

### 6. Distinguishing land from buildings is particularly relevant for Welsh policy goals

The case studies show that many countries value land separately from buildings even when they tax them jointly. This has clear relevance for Wales:

- Land-only valuation could support land value capture, infrastructure planning, compensation, public land transactions, or reform of non-domestic rates.
- It allows policy to target land-value uplift, not improvements.
- It provides a consistent base for cross-boundary planning and investment decisions across Welsh local authorities.

A system that cleanly separates land from improvements would therefore enhance Wales's ability to pursue broader policy reforms (e.g., land-value taxation, infrastructure levy, planning gain alternatives).

### 7. Wales should consider how land valuation ties into broader strategic objectives

Several international systems show how land-valuation design supports broader policy goals:

- Montana protects agricultural land from speculative pressure via use-value.
- Hamburg uses location factors to influence housing affordability.
- Luxembourg introduces higher charges on vacant/undeveloped land to stimulate development.

- Japan relies on stable unit values to facilitate inheritance planning.

Wales may wish to consider using land valuation to support:

- Housing delivery and land-release incentives,
- Fair and transparent local fiscal reform,
- Rural land stewardship,
- Decarbonisation and nature targets,
- Improved planning consistency and certainty.

## STRAND B FINDINGS: STRUCTURE-ONLY HEDONIC MODEL WITH MSOA FIXED EFFECTS

To explore whether a practical formula-based valuation approach could be derived from Welsh data, we estimated a hedonic regression containing:

- only structure-related variables (e.g. adjusted freehold parcel area, construction age, property type, floor level), and
- MSOA fixed effects capturing all locational influences.

This mirrors the approach used in ap Gwilym et al. (2020) and is statistically analogous to "standardised land values" used in several international models identified in Strand A.

**Table 3.3.16: Model fit and performance by property type**

| Property type | Number of Transactions | RMSE (£) | $R^2$ |
|---|---|---|---|
| **Residential** | 1,105,650 | 358,685 | 0.121 |
| **Unknown or Not Stated** | 163,374 | 1,168,386 | 0.021 |
| **Retail and Commercial** | 7,037 | 5,034,571 | 0.008 |
| **Food/Drink and Hospitality** | 3,112 | 2,368,790 | 0.008 |
| **Offices** | 2,032 | 6,153,012 | 0.004 |
| **Residential Institutions** | 1,473 | 9,168,555 | -0.005 |
| **Industrial and Warehousing** | 713 | 14,402,693 | -0.007 |
| **Healthcare** | 392 | 4,406,922 | 0.000 |
| **Community, Education and Public Buildings** | 264 | 4,404,857 | 0.000 |
| **Assembly and Leisure** | 140 | 4,368,928 | -0.007 |
| **TOTAL** | 1,284,218 | 309,914 | 0.154 |

## MODEL PERFORMANCE

The overall explanatory power of the model is:

- $R^2 = 0.154$ which is materially weaker than the full hedonic model in Lot 1.

- RMSE for the overall model is £309,914, which is again materially worse than either Lots 1 or 2.

R2 and RMSE by property type and local authority area are presented in Tables 3.3.16 and 3.3.17 respectively.

While the model still performs reasonably well for such a simplified specification, this confirms that removing detailed land-related variables inevitably reduces explanatory power.

**Table 3.3.17: Model fit and performance by local authority**

| Local authority | Number of Transactions | RMSE (£) | $R^2$ |
|---|---|---|---|
| **Blaenau Gwent** | 24,524 | 287,879 | 0.044 |
| **Bridgend** | 61,694 | 749,445 | 0.023 |
| **Caerphilly** | 69,629 | 894,512 | 0.014 |
| **Cardiff** | 162,546 | 1,403,716 | 0.015 |
| **Carmarthenshire** | 71,328 | 908,673 | 0.012 |
| **Ceredigion** | 23,480 | 370,426 | 0.035 |
| **Conwy** | 57,475 | 399,966 | 0.055 |
| **Denbighshire** | 43,987 | 442,998 | 0.039 |
| **Flintshire** | 61,545 | 833,796 | 0.016 |
| **Gwynedd** | 43,187 | 528,505 | 0.033 |
| **Isle of Anglesey** | 26,401 | 409,547 | 0.052 |
| **Merthyr Tydfil** | 21,147 | 594,326 | 0.019 |
| **Monmouthshire** | 41,592 | 1,627,483 | 0.01 |
| **Neath Port Talbot** | 56,701 | 446,618 | 0.031 |
| **Newport** | 68,681 | 683,321 | 0.03 |
| **Pembrokeshire** | 49,259 | 700,275 | 0.02 |
| **Powys** | 45,111 | 371,202 | 0.06 |
| **Rhondda Cynon Taf** | 103,331 | 259,186 | 0.083 |
| **Swansea** | 102,576 | 1,099,402 | 0.014 |
| **Torfaen** | 34,426 | 564,417 | 0.028 |
| **Vale of Glamorgan** | 67,209 | 646,467 | 0.056 |
| **Wrexham** | 48,389 | 521,182 | 0.04 |
| **TOTAL** | 1,284,218 | 309,914 | 0.154 |

## COEFFICIENT INTERPRETATION

The structure-related coefficients behave sensibly. Adjusted Freehold Parcel Area is, unsurprisingly, the dominant driver of total price. Structure attributes (e.g. construction age, property type, floor level) are statistically significant but of modest magnitude. Signs and relative magnitudes are broadly consistent with the patterns observed in Lot 1.

The notable exception is the intercept term, which is large, statistically significant, and sharply different from the Lot 1 model, where the intercept was not significantly different from zero.

This contrast highlights that MSOA dummies are a weak substitute for explicit land attributes. In Lot 1, land characteristics enter the model directly; in Lot 3, the intercept is forced to absorb much of their effect.

## MSOA EFFECTS

The raw MSOA coefficients exhibit plausible spatial patterns, consistent with broad expectations of land values across Wales. They display reasonable differentiation between higher-value and lower-value locations, but poor calibration in absolute terms, because the structure-only model lacks the information needed to place land values on the correct monetary scale.

## CALIBRATION OF MSOA EFFECTS

The raw MSOA coefficients provide credible relative differences between locations, but the absolute levels are not correctly centred. This is a direct consequence of the simplified model specification which lacks explicit spatial variables. With no reliable bare-land transaction data in Wales to anchor the level, the model is unable to identify the absolute price of land.

This is an especially acute issue in Wales because bare-land sales are extremely rare, and the few that do exist are often non-standard, tax-advantaged or highly idiosyncratic parcels that do not reflect market-wide land values.

To address this, we calibrated the Lot 3 outputs using the results from Lot 2, which provides the most comprehensive modelling of structure–land separation, with the clearest internal consistency across Wales.

Accordingly, we adjusted the intercept term of the Lot 3 model so that the mean land value in Lot 3 matches the mean land value implied by the Lot 2 estimates, while preserving the relative differences across MSOAs produced by the Lot 3 regression.

This produces a calibrated set of MSOA-level land values suitable for generating formula-based valuations.

## LOT 3 LAND VALUE ESTIMATES

Land value estimates based on the calibrated MSOA coefficients are presented in the tables and maps below.

In order to estimate land values, the parameters on structure attributes were set equal to zero. The only remaining variables are Adjusted Freehold Parcel Area, MSOA and the intercept term. The intercept term is adjusted to calibrate the model. In this case, we have calibrated the model to fit the mean land value in the Lot 2 residential model, as reported in Table 3.2.9. The resulting summary statistics for land valuation in the Transactions Database are shown in Table 3.3.18.

**Table 3.3.18: Lot 3 land value estimates for the Transactions Database**

| Minimum | Median | Mean | Max |
|---|---|---|---|
| £3,658 | £175,458 | £178,264 | £25,099,491 |

We also used the model to generate out-of-sample land-value estimates for every parcel in both the National Land Parcel Database and the LSOA Land Parcel Database. The results for the latter are summarised in Table 3.3.19.

**Table 3.3.19 Lot 3 land value estimates for the LSOA Land Parcel Database**

| LSOA | Number of observations | Median land value | Mean land value |
|---|---|---|---|
| **Trawsfynydd** | 5,652 | £111,438 | £143,527 |
| **Broughton North East** | 1,639 | £173,639 | £174,453 |
| **Knighton 2** | 1,280 | £162,358 | £163,284 |
| **Aberystwyth Rheidol 1** | 616 | £223,434 | £223,458 |
| **Scleddau** | 4,105 | £170,848 | £174,442 |
| **Porthcawl East Central 2** | 1,038 | £233,344 | £233,468 |
| **Rhigos** | 3,961 | £128,705 | £138,390 |
| **Usk 2** | 1,070 | £257,112 | £257,408 |
| **Cathays 12** | 1,119 | £281,194 | £281,404 |

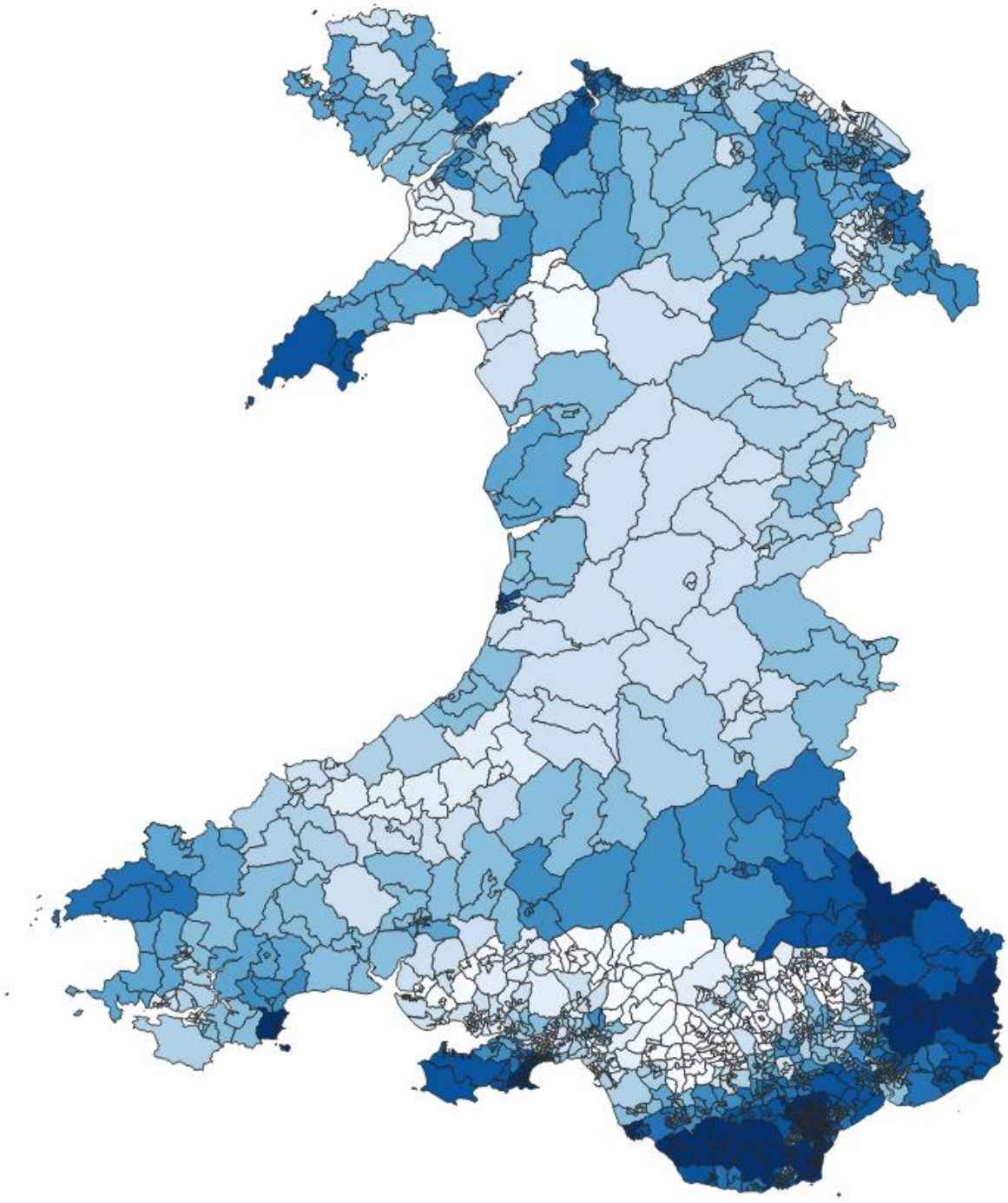
Figure 3.3.1 illustrates the average land value per parcel in each LSOA, based on parcels in the National Land Parcel Database. Figure 3.3.2 illustrates land values per $m^2$ for each of the polygons in the LSOA Land Parcel Database. Clicking on the images will take you to an interactive, scrollable version of each map.

These maps show a broad regional structure of land values consistent with Lots 1 and 2, though there are some differences in detail. The lowest land values are found in the south Wales valleys, and also in Ffestiniog, Dyffryn Nantlle, Rhyl and

Swansea. The highest land values are found in parts of Cardiff, the Vale of Glamorgan, Monmouthshire, southern Swansea/Mumbles, and Tenby.

Differences in detail are expected. MSOA fixed effects impose a coarse spatial structure, smoothing out localised effects that the Lot 1 and Lot 2 models explicitly capture.

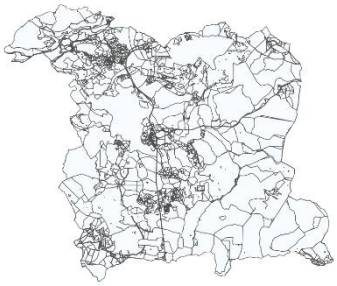
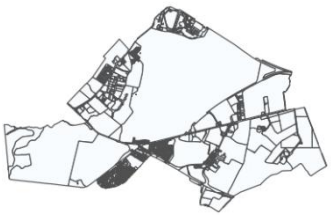
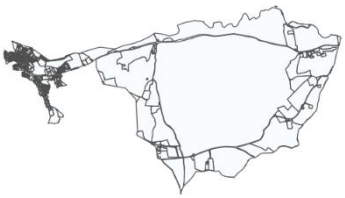
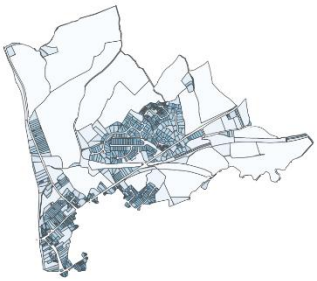
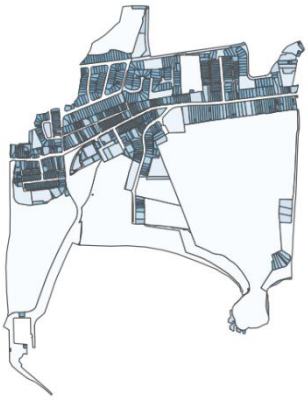
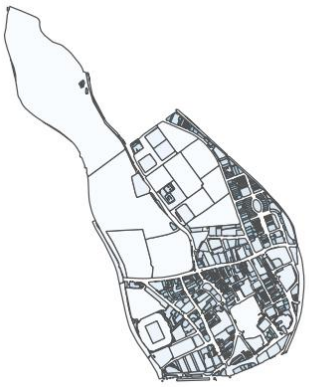
**Figure 3.3.1: Lot 3 average land values per parcel in the National Land Parcel Database by LSOA**



Darker shades of blue represent higher land values per parcel.

Please click on the map to access an interactive, scrollable version.

**Figure 3.3.2: Lot 3 land values per m$^2$ at the parcel level**

| | | |
|---|---|---|
| Trawsfynydd:  | Broughton North East:  | Knighton 2:  |
| Aberystwyth Rheidol 1:  | Rhigos:  | Usk 2:  |
| Scleddau:  | Porthcawl East Central 2:  | Cathays 12:  |

Darker shades of blue represent higher land values per parcel.

Please click on the map to access an interactive, scrollable version.

## INTERPRETATION AND IMPLICATIONS

The Lot 3 results demonstrate that a statistically derived, MSOA-based formula approach is feasible, in the sense that it produces plausible spatial variation, it is potentially easier to explain, and it is grounded in Welsh data.

However, it relies critically on calibration from more sophisticated models, or reliance on bare-land sales data that is extremely limited in the Welsh context. On its own, the structure-only MSOA model cannot recover absolute land values.

The data challenges are essentially identical to those in Lots 1 and 2, because structure data (the true bottleneck) is still required. Spatial data is not a binding constraint. Hence, the most important variables in the Lot 1 and Lot 2 models cannot simply be “replaced” by MSOA.

There is no strong technical justification for pursuing a formula-based system of this type. Technically, it adds an extra layer of indirection. Policymakers want a simple formula because complex models are hard to communicate. However, the simple formula can only be produced by using those complex models for calibration.

Nevertheless, political considerations may favour simple, stable, easily communicated formulae. If policymakers wish to adopt a formula-based system for reasons of public comprehension or administrative convenience, Lot 3 shows that such a system is possible but it must be supported by periodic re-estimation and calibration using full hedonic or ML models.

## SUMMARY

Taken together, the two strands of Lot 3 show that formula-based land valuation systems fall into three distinct families internationally, each with very different implications for Wales. The first consists of models that apply simple statutory formulae based on land area, sometimes with categorical adjustments, but without any reference to land value as understood in economic terms. These systems achieve simplicity by design, but they do not attempt to measure land value and therefore offer little analytical or policy insight for Wales beyond illustrating how minimalist some jurisdictions are willing to be.

At the other extreme are systems that rely on full mass appraisal, usually using hedonic, comparable-sales, or other market-based techniques. These do not use formulas in any meaningful sense; they rely on the same kinds of statistical infrastructure and modelling practices as the approaches developed in Lots 1 and 2. For Wales, these systems are already represented within this report: the full hedonic framework of Lot 1 and the machine-learning models of Lot 2 are direct equivalents

of the international mass-appraisal models used in places such as Denmark, the Netherlands, Australia, and South Africa.

Between these two poles lie systems that apply standardised unit land values, typically defined for zones, neighbourhoods, road segments, or cadastral blocks. Such systems are formula-based in the sense that land value is produced mechanically by multiplying parcel area by a unit value, but the unit values themselves are produced through administrative assessment, expert panels, or reference to confidential valuation rolls. The underlying methods are therefore not directly visible, and the systems cannot be replicated from public data alone.

Strand B explored whether a Welsh equivalent of this middle category could be created statistically, using a structure-only hedonic model with MSOA fixed effects to generate a set of implicit, standardised location values. This exercise shows that it is indeed possible to construct such a system, and the resulting MSOA effects produce spatial patterns broadly consistent with those in Lots 1 and 2. However, the model cannot recover credible absolute land values without external calibration, because removing explicit land-related variables leaves the intercept and MSOA coefficients poorly centred. In Wales this limitation is especially acute: bare-land transactions—which in principle could provide an external anchor—are extremely scarce, and those that do exist are typically atypical, tax-advantaged or idiosyncratic plots whose prices cannot be relied upon to represent market-wide land values.

As a result, any formula-based valuation system in Wales must ultimately be calibrated against a richer model, such as those developed in Lots 1 or 2. This means that, although the final output may appear simpler and more transparent to non-specialists, it depends analytically on the very modelling frameworks it is meant to replace. This creates a fundamental tension: there is no strong technical justification for adopting formula-based approaches of this kind, since they offer no reduction in data requirements and no simplification of the underlying analytical task. Nonetheless, political considerations – particularly around transparency, communication, and administrative stability – may lead policymakers to prefer formula-based systems despite their technical limitations.

## LOT 5: INNOVATIVE OR EXPERIMENTAL APPROACHES

In this section we present the results of Lot 5. We begin by summarising the data collected and compare the decisions from each of the two treatments in order to examine how different incentive schemes influence residents' choice of land valuation methodologies. We then move to examine how individual characteristics correlate with decisions. Whenever we perform a statistical test, we use the 5% statistical level to establish statistical significance. We present the p value associated with the statistic in brackets. All statistical tests are two sided, unless otherwise stated.

### SUMMARY STATISTICS

A total of 201 residents completed the experiment. Table 3.4.1 summarizes the data collection, presenting the number of observations for each treatment, and the number of residents selecting each of the LVMs.

**Table 3.4.1: Chosen land valuation methodologies by treatment**

| Treatment | Observations | Hedonic pricing | Machine learning | Equation based |
|---|---|---|---|---|
| ***Low*** | 104 | 46 | 40 | 18 |
| ***High*** | 97 | 41 | 32 | 24 |

Due to the nature of the randomization procedure, slightly more residents were randomized into the *Low* versus *High* treatment. This has no impact on the statistical analysis.

Table 3.4.2 summarizes the residents that took part, presenting information on the accuracy of the data presented to them, the percentage of them that are home-owners and the number of comprehension questions they got correct in Stage 1.

**Table 3.4.2: Chosen land valuation methodologies by treatment**

| Treatment | Obs. | Accuracy | Home-owners | Comprehension |
|---|---|---|---|---|
| ***Low*** | 104 | 7.79<br>(2.11) | 58% | 3.37<br>(0.791) |
| ***High*** | 97 | 7.97<br>(2.06) | 62% | 3.41<br>(0.732) |

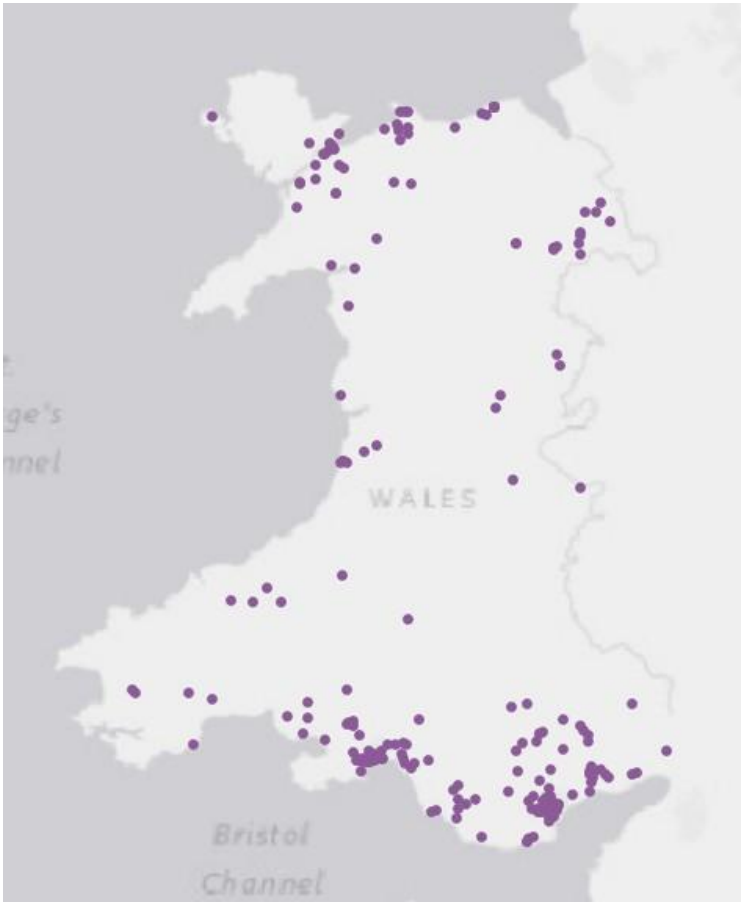
Note: Accuracy 1-10 (10 being most accurate). Home-owners as a percentage of respondents. Comprehension questions correct, 1-4 (4 being all correct). Standard deviations in brackets.

From Table 3.4.2, there are only small differences between the two treatments, and there are no statistical differences in the accuracy (p=0.50, Robust Rank order Test) or the number of comprehension questions that residents got correct (p=0.88, Robust Rank order Test) between treatments. We do however find that there are

significant differences in resident type between treatments (p=0.02, Fisher's exact test). This is not a consequence of there being more home-owners in the *High* treatment in comparison to the *Low* treatment, but instead because a relatively high number of residents reported, "prefer not to say" in the *High* treatment. There is no significant difference between the proportion of home-owners or renters between treatments (p=0.11, Fisher's Exact test).

Figure 3.4.1 presents a map of the spatial distribution of the residents that took part: each point on the map of Wales represents a single resident, but there are a number of overlaps. As can be seen, residents from the North, South, East and Wales participated, including both urban and rural areas.

**Figure 3.4.1: Spatial distribution of the 201 residents that took part in the experiment**




We divide Wales into four regions based on the *Growth Deals* signed between 2016 and 2022. Those four regions are the *North Wales Growth Deal*, the *Mid Wales Growth Deal*, the *Cardiff Capital Region* and the *Swansea Bay City Region*. Table 3.4.3 presents the number of residents in each region in each treatment, based on their reported postcode.

**Table 3.4.3: Spatial distribution of residents**

| Treatment | Obs. | Cardiff Capital | Mid Wales | North Wales | Swansea Bay |
|---|---|---|---|---|---|
| ***Low*** | 104 | 43 | 7 | 29 | 25 |
| ***High*** | 97 | 35 | 14 | 27 | 21 |

We report no statistical differences in the location of participants between the treatments (p=0.349, Fisher's exact test).

As we only find that residency type differs between treatment, a consequence of more participants reporting "prefer not to say", we take this as evidence that the types of residents assigned to each treatment is well balanced, and that our randomization procedure was successful.

## DOES THE INCENTIVE TO MINIMIZE OR MAXIMIZE LAND VALUE INFLUENCE CHOICES?

Table 3.4.1 presents the number of residents that selected each of the LVMs in each of the two treatments. As can be seen, decisions in both the *Low* and *High* treatments are remarkably similar and in both the treatments residents select *Hedonic Pricing* most frequently, followed by *Machine Learning* and then *Equation Based*. This leads to our first observation.

### Observation 1. The hedonic pricing method is the most frequently chosen land valuation methodology.

Comparing the proportions of each LVM selected in each treatment, we observe no statistical differences. Residents made near identical choices regardless of the incentives they faced. A parametric statistical analysis, where we control for potential differences between the residents in each treatment, shows that there are no differences in residents' choices between the treatments, no matter the controls used. This analysis is provided in Appendix C. This leads us to our second observation.

### Observation 2. The incentive to select the land valuation methodology that minimizes or maximizes land value has no influence on residents' choices.

One interpretation of Observations 1 and 2 is that the *Hedonic Pricing* LVM is the most preferred method. As *Hedonic Pricing* was selected most frequently regardless of treatment, it may be that there is something about the *Hedonic Pricing* method that residents inherently prefer to the other two LVMs, and that the incentives we offered them were not enough to get them to decide otherwise. However, it seems unlikely that residents completing the experiment would be willing to forgo between £5 and £10 pounds simply to select the *Hedonic Pricing* method if they believed one

of the other LVMs was more profitable for them. An alternative interpretation of Observation 1 and 2 may be that residents do not know which of the LVMs will produce the highest or lowest land valuations, which is why we observe no differences between the treatments. It may be that this is a response to the descriptions, or labelling, of the methodologies we provided.

## ARE CHOICES A RESULT OF HOW WELL RESIDENTS UNDERSTAND LAND VALUATION?

In Stage 1 of the experiment, subjects were required to answer four comprehension questions after watching a video that explained land valuations. The number of correctly answered comprehension questions likely reveals how well a resident understood the video and its contents. This understanding may be an important correlate with the decision to select a specific LVM.

**Observation 3. There is no correlation between comprehension of land valuation and the choice of land valuation methodology.**

Using parametric analysis (reported in the Appendix) we find no evidence that the number of correctly answered comprehension questions correlates with the choice of LVM in either treatment. Observation 3 is reassuring, as it suggests that the LVM choices we observe are not a consequence of resident misunderstanding land valuation.

## DO DIFFERENT RESIDENTS MAKE DIFFERENT CHOICES?

As can be seen from Figure 3.4.1, our experiment has a sample of residents from across Wales, with people reporting addresses in the North, South, East and West. The location of residents is balanced across treatments, as highlighted by Table 3.4.3. We also have residents that are home-owners, tenants and some who prefer not say, with slight differences as highlighted in Table 3.4.2. It may be that the selection of LVM correlates with some of these characteristics. In this section, we explore this further.

### NORTH WALES, MID WALES, SWANSEA BAY AND CARDIFF CITY

We begin by examining if residents from four regions, as defined by the *Growth Deals* signed between 2016 and 2022, make different choices. Those four regions are the *North Wales Growth Deal*, the *Mid Wales Growth Deal*, the *Cardiff Capital Region* and the *Swansea Bay City Region*. These regions differ along a number of characteristics. For example: North and Mid Wales are both very rural, whilst the Cardiff Capital Region is urban. North Wales has high proportion of Welsh speakers. It may be that residents from these differ in their preferences, or knowledge, about land valuations, or in their perception of specific methodologies. The statistical analysis presented in Appendix C, Table C1, reveals there are no significant

differences resulting from the region a resident is located – choices are near identical, regardless of region.

### HOMEOWNERS, TENANTS AND OTHERS

A second important consideration is the type of resident we are considering: home-owners and tenants may make different choices; those who "prefer not to say" may also differ in their choices. This could be true for a variety of reasons: home-owners may wish to select the LVM that maximizes the value of their land, regardless of the experimental incentives, whereas tenants might wish to minimize it. The statistical analysis presented in Appendix C, Table C1, supports the conclusion that residency type does not correlate with LVM choice. This is true, regardless of treatment, and residency type is never a statistically significant correlate of LVM choice.

We report little evidence that choice of LVM correlates with our residents' individual characteristics and circumstances. This leads to our fourth observation.

**Observation 4. Residents' characteristics and circumstances do not correlate with their choice of land valuation methodology.**

## CONCLUSION

We conducted an online economic experiment to evaluate how 201 Welsh residents evaluate three different land valuation methodologies – Hedonic Pricing, Machine Learning and Equation Based modelling. Residents were asked to select which of these methodologies they would like to be used to value the land associated with their property, and they were incentivised to select the methodology that either (1) maximized the value of their land or (2) minimized the value of their land.

We observe that Welsh residents (1) are most likely to select the Hedonic Pricing method and (2) that this choice does not vary with the incentives they face. We also find evidence that (3) this is not a consequence of residents that are more able to comprehend land valuation selecting this methodology. Finally, we observe that (4) the residents' individual circumstances do not correlate with their decisions.

Explaining why residents' choice of LVM does not change with the incentives they face – either to maximize or minimize the value of their land – is unclear. One explanation is that they have an inherent preference for the Hedonic Pricing method. An alternative, and perhaps more likely explanation, is that they are unsure how the different methodologies will value their land and therefore how they will impact their earnings in the experiment. Another alternative explanation may be that the Hedonic Pricing method being most popular is a response to the wording used for the Machine Learning or Equation Based methods, rather than because the methodology is preferred *per se*. In this relatively small exploratory study, we are unable to distinguish between these competing explanations.

Future work should focus on disentangling why specific land valuation methodologies are favoured by residents over others – and to what extent residents' choices can be manipulated by how the methodologies are framed and described.

# 4. COMPARISON OF APPROACHES

This section synthesises the outputs of the four lots and sets out how the different approaches compare in terms of methodology, performance, interpretability, data requirements and operational feasibility. The four methods are not alternatives in the sense of being mutually exclusive; each reflects different modelling philosophies and offers different kinds of insight into land valuation in Wales. The objective here is therefore to compare their properties, highlight the trade-offs that arise, and clarify what each approach can and cannot deliver.

Land valuation is inherently difficult because, unlike property values, there is no observable “true” land value against which estimates can be validated. All four approaches rely on the same underlying transaction prices and parcel/structure information, and all produce land values by setting either land or structure-related attributes to zero. As a result, accuracy can only be assessed for predicted total property prices, not land values themselves, and differences in land-value estimates reflect modelling assumptions rather than objective benchmarks.

## OVERVIEW

The four lots adopt distinct approaches to land valuation:

- **Lot 1** uses hedonic regression to decompose property prices into land and structure components using an explicitly specified functional form.
- **Lot 2** uses machine-learning models, which detect interactions and non-linearities algorithmically rather than through human-imposed functional form.
- **Lot 3** explores formula-based valuations similar to those used internationally, relying on simplified modelling structures and requiring external calibration.
- **Lot 5** uses experimental methods to study how people respond to different valuation approaches and the extent to which they trust or prefer them.

While Lots 1–3 all follow the same broad logic – estimate a model of total property values, then derive land values by removing structure variables – they differ in where the **modelling discretion** lies. In hedonic modelling, decisions about model specification, transformations and interactions are made explicitly by the modeller. In machine-learning models, many of these decisions are embedded inside the algorithm. Formula-based models, by contrast, intentionally minimise the number of variables included, but this simplicity creates additional challenges in anchoring land values on a meaningful numerical scale.

## SUMMARY COMPARISON

Table 4.1 provides a high-level comparison of the four approaches. It is intended as a quick reference for the narrative that follows.

**Table 4.1: Summary of the four approaches**

| Dimension | Lot 1: Hedonic Regression | Lot 2: Machine Learning | Lot 3: Formula-Based | Lot 5: Experimental Preferences |
|---|---|---|---|---|
| **Modelling approach** | Explicit functional form; modeller chooses transformations, interactions and variables | Algorithmic discovery of interactions and non-linearities | Structure-only regression with MSOA fixed effects | Behavioural experiment with real incentives |
| **Decomposition of land & structure** | Yes – achieved by setting structure variables to zero | Yes – achieved by setting structure variables to zero | Yes – but needs external calibration for intercept weight, $\theta$ | Not applicable |
| **Transparency and interpretability** | Moderate – structure is explicit, but many variables and decisions | Moderate–low – structure less explicit; some choices automated | High in form, but calibration step adds opacity | High (focus on understanding, not valuation) |
| **Accuracy (property values only)** | Moderate | Highest (particularly XGBoost) | Lowest | Not applicable |
| **Primary purpose** | A benchmark decomposition of property prices into land and structure values | Maximise predictive accuracy | Create simple, transparent valuation rules | Assess public trust and preference |
| **Modelling discretion** | High – modeller defines structure of model | Moderate – many choices embedded within the model | Low – intentionally simplified specification | Not applicable |
| **Accuracy** | Moderate | Highest | Lowest | N/A |
| **Data requirements** | High | High | High | Moderate |
| **Operational feasibility** | Feasible but accuracy depends on data quality | Feasible; same data constraints as Lot 1 | Least feasible due to external calibration | Feasible for engagement, not valuation |
| **Public acceptability** | Highest | Moderate | Lowest | Hedonic model most preferred |
| **Best suited to…** | Contexts where explicit control of model structure and decomposition is important | Contexts where accuracy is prioritised and algorithmic discovery is valuable | Contexts prioritising simplicity where external calibration is acceptable | Understanding public perceptions and trust in valuation approaches |

# TECHNICAL PERFORMANCE ACROSS LOTS

## HEDONIC REGRESSION (LOT 1)

The hedonic model provides a structured, explicitly specified link between structural, locational and environmental characteristics and total property values. The model achieves an $R^2$ of 0.441 across Wales, with stronger performance in the residential segment and weaker performance in more heterogeneous property categories. These performance levels are typical of hedonic models in diverse housing markets and indicate that the model captures a reasonable proportion of systematic variation in transaction prices.

It is important to recognise that the hedonic model's explanatory variables are numerous, and many exert small but statistically significant effects. In practice, this means that the model does not provide a simple narrative about "the main drivers of land value". Instead, it reflects the reality that property prices are shaped by a very wide range of factors, none of which can be cleanly isolated as dominant.

## MACHINE LEARNING (LOT 2)

Machine-learning models achieve markedly higher predictive performance. After removing outliers and applying log transformations to skewed variables, the best LASSO-based models deliver cross-validated RMSE values around £85,000. More flexible methods, such as neural networks and XGBoost, perform even better, with XGBoost achieving an RMSE of £68,831.

These gains arise because machine-learning models automate the identification of interactions and non-linear relationships that hedonic models can only include if explicitly specified. However, while their predictive performance is strong, the internal structure of such models is complex, with many modelling decisions (splitting rules, non-linear transformations, regularisation) encoded within the algorithms themselves. As a result, machine-learning models are not inherently more transparent or less discretionary. Rather, they relocate some modelling discretion into algorithmic design rather than explicit functional form.

Furthermore, while these models can generate derived estimates of land value by holding structure variables at zero, their flexibility means that such decompositions are not unique or incontrovertible.

## FORMULA-BASED MODELS (LOT 3)

Formula-based models intentionally remove most land-related variables and captures location effects through limited variables (in our case, through MSOA fixed effects). This yields relatively low explanatory power ($R^2 = 0.154$), which is consistent

with the simplified specification. The simplicity, however, also means that the model cannot determine the absolute level of land values internally. Without bare-land transactions in Wales, the intercept cannot be pinned down, and external calibration (using Lot 2 in this project) is required.

Despite their lower accuracy, these models remain valuable as parsimonious, transparent representations of spatial value patterns. International experience shows that such systems can be effective where simplicity, predictability and administrative ease are prioritised over fine-grained accuracy. However, the reliance on calibration distinguishes Lot 3 from Lots 1 and 2. While it offers simplicity of structure, that simplicity is achieved by deferring part of the valuation to a richer model.

## TRANSPARENCY, MODELLING DISCRETION AND INTERPRETABILITY

A key insight from this project is that none of the approaches is straightforward or "transparent" in a simple sense. All contain significant complexity:

- Lot 1 requires many modelling decisions (variable selection, transformations, interactions).
- Lot 2 embeds many such decisions inside algorithms.
- Lot 3 requires an external calibration step that is itself opaque.

The primary distinction is therefore not transparency but the locus of modelling discretion.

- Hedonic regression places discretion with the analyst.
- Machine-learning models embed discretion within algorithms.
- Formula-based approaches place modelling discretion with the architects of the formula itself, particularly through the way calibration choices are embedded in the structure.

## OPERATIONAL FEASIBILITY

All four lots were completed successfully, demonstrating that each approach is operationally feasible given current data. The primary limitation for all methods is the quality and completeness of underlying datasets, particularly for non-residential characteristics, planning data and amenity quality.

Within these constraints:

- Lots 1 and 2 rely on the same core data and are equally feasible to implement in principle, though neither achieves accuracy that would be acceptable for some statutory or fiscal applications.
- Lot 3 appears simpler but is least directly feasible, because it cannot generate land values without external calibration from another model.
- Lot 5 is operationally straightforward, as it concerns behavioural preferences rather than valuation.

## PUBLIC PREFERENCES AND ACCEPTABILITY

The experimental evidence from Lot 5 shows that participants most preferred the hedonic approach, followed by the machine-learning and formula-based methods. Importantly, these preferences were stable across incentive treatments, comprehension levels and participant characteristics. This suggests that public trust is more strongly influenced by the perceived legitimacy and logical structure of the method than by expected financial outcomes. This insight is critical for any future implementation of land valuation, where acceptability may depend as much on perceived fairness as on technical performance.

## SYNTHESIS

The four Lots highlight that land valuation involves a series of unavoidable trade-offs. Hedonic modelling provides explicit structure but requires many judgement calls. Machine-learning methods offer greater accuracy but embed complexity within algorithms. Formula-based models offer superficial simplicity but require calibration and sacrifice precision. Behavioural evidence underscores the importance of approaches that people can understand and trust.

The choice between methods therefore depends on the policy context rather than any intrinsic superiority of one approach. All three modelling approaches can produce internally coherent land estimates given the data available in Wales, but none can be externally validated against true land values. The appropriate approach will depend on the purpose of the valuation exercise, and on how Welsh Government wishes to balance accuracy, interpretability, discretion, administrative simplicity and public acceptability.

# 5. CONCLUSIONS

This project set out to test, compare and understand a range of methodologies for estimating land values in Wales. Drawing on a uniquely rich collection of administrative, geospatial and constructed datasets, and applying a suite of analytical and behavioural approaches across four lots, the project provides the clearest evidence to date on what is, and is not, possible in the Welsh context. The conclusions reflect not only the statistical results but also the conceptual, institutional and behavioural insights that emerged over the course of the work.

## Conclusion 1: We can estimate land values in Wales, but only within the limits of its fragmented data environment

A central conclusion of this project is that it is possible to construct coherent, internally consistent estimates of land value across Wales, but the accuracy and reliability of these estimates are fundamentally constrained by the underlying data infrastructure.

Although we have assembled comprehensive land-valuation datasets, drawing together transactions, EPCs, INSPIRE polygons, spatial amenities, environmental indicators and socio-economic variables, the data we rely on were not created for land-valuation purposes. They vary in coverage, resolution, consistency and purpose, and lack common identifiers that would allow seamless linkage.

Problems with overlapping INSPIRE polygons, properties without UPRNs, bundled transactions, limited planning data, lack of amenity-quality measures, and missing data throughout all introduce uncertainty that cannot be resolved through statistical methods alone. These limitations are structural and reflect the absence of a unified Welsh cadastral institution with responsibility for maintaining authoritative parcel-property-attribute linkages.

## Conclusion 2: Despite these constraints, the modelling results represent a step change in the Welsh evidence base

The modelling work demonstrates significant progress over previous studies. The Lot 1 hedonic model achieves a higher explanatory power than ap Gwilym et al. (2020), and Lot 2 machine-learning models achieve higher predictive accuracy still.

A particularly important result is that the intercept term in the Lot 1 model is not statistically different from zero. This indicates that the combined dataset captures a sufficient share of land-related variation that the model does not rely on an arbitrary constant to anchor valuations. This is strong evidence that the integrated dataset functions as an effective proxy for the underlying determinants of land value.

### Conclusion 3: “Land value” is not an observable quantity, and decomposing property values requires modelling assumptions

Another key conclusion is philosophical.

Unlike property prices, land value does not exist as an observable market quantity in Wales. Land is almost always transacted with structures, and rarely in a way that reveals its stand-alone value.

All methods in this project – hedonic, machine-learning and formula-based – therefore rely on assumptions to separate land from structures. The separability assumed in the Lot 1 and Lot 2 decomposition methods is a modelling choice rather than an empirical fact. Indeed, the superior performance of non-separable models such as XGBoost provides empirical support for the idea that location and structure are deeply complementary and cannot be cleanly disentangled through any single method.

Moreover, land value means different things in different contexts. For insurers, it may mean the residual value of a site absent structures. For a land-value tax, it may represent the socially attributable value of location. For planners or valuers, it may reflect development potential.

There is therefore no single, universal definition of “land value”, and any operational valuation system must be explicit about what conceptual definition it aims to approximate.

### Conclusion 4: The strongest differences between approaches relate to the location of modelling discretion, not transparency or simplicity

None of the modelling approaches is transparent or simple in any absolute sense. Each contains significant complexity:

- Hedonic regression requires explicit decisions about variable construction, transformations, interactions, and model specification.
- Machine-learning models embed many such decisions within algorithmic structures and hyperparameters.
- Formula-based models minimise the number of variables but shift discretion into the choice of calibration and the assignment of baseline values.

The key distinction is therefore where the judgement sits: with the modeller (Lot 1), the algorithm (Lot 2), or the architect of the formula (Lot 3). No approach eliminates judgement; each simply distributes it differently.

### Conclusion 5: The three modelling approaches produce broadly consistent spatial patterns across Wales

Despite their differences, Lots 1–3 all produce recognisable and consistent spatial measurements of land value. Lower land values appear systematically in the South Wales Valleys and certain post-industrial or peripheral communities. Higher land values cluster around Cardiff, Monmouthshire, parts of Swansea, and specific coastal or amenity-rich areas.

This robustness across modelling paradigms suggests that the spatial geography of land value in Wales is deeply rooted in observable features of the built and natural environment, even if the absolute levels of valuation differ across methods.

### Conclusion 6: Lot 5 shows that public preferences are shaped by perceived understandability, not incentives

The behavioural experiment demonstrates that participants consistently preferred the hedonic model, followed by machine learning and then formula-based approaches. Crucially, these preferences do not change under different financial incentives.

This is not surprising given that participants had no reliable way of predicting which model would yield higher or lower valuations for their own parcels. The experiment therefore reveals preferences based on perceived intelligibility and perceived legitimacy, rather than strategic behaviour. This reinforces the importance of approaches that people feel they can understand and that appear grounded in recognisable logic, even if those approaches are not the most statistically accurate.

## OVERALL SYNTHESIS

This project shows that land valuation in Wales is both possible and inherently constrained. The modelling work demonstrates clear progress and strong internal coherence, but the accuracy of estimated land values is limited by data that is incomplete, inconsistent and not designed for the purpose. The philosophical nature of land value further limits what can be concluded: any estimate necessarily reflects modelling assumptions rather than a directly observable quantity.

At the same time, the project reveals stable spatial patterns across methods and provides new insight into how people perceive the legitimacy of valuation approaches. These findings form a robust foundation for future discussions about land valuation in Wales, but they also highlight that meaningful progress will require institutional and data-infrastructure reforms. The choice of valuation method will ultimately depend on the specific policy context and on how Welsh Government wishes to balance accuracy, interpretability, discretion, administrative feasibility and public acceptability.

# 6. FUTURE CONSIDERATIONS

The findings from this project point to several issues that Welsh Government may wish to consider when reflecting on the future of land-value analysis and its potential applications in Wales. These considerations do not prescribe specific actions but highlight areas where strategic reflection will be necessary if Wales is to make fuller use of land valuation methodologies in the future.

## THE NEED FOR A COHERENT, AUTHORITATIVE CADASTRAL DATA INFRASTRUCTURE

A central theme across the four lots is that the greatest constraint on land-value modelling is not the choice of method but the underlying data environment. Although this project brought together the widest collection of relevant datasets possible, these sources were developed for disparate administrative purposes, lack consistent identifiers, and frequently contain gaps, overlaps or inconsistencies. The difficulties encountered in matching INSPIRE polygons to UPRNs, resolving overlapping parcel boundaries, handling bundled transactions in PPD, and integrating EPC and spatial attributes all reflect a deeper structural issue: Wales does not currently possess a unified cadastral data framework.

This echoes a key conclusion of ap Gwilym et al. (2020), which recommended establishing a Welsh cadastral database under a single agency. The experience of this project provides further empirical support for revisiting that earlier recommendation. International examples show that such investments can deliver substantial long-term value. For instance, the Montana cadastral project (completed in 2003) cost approximately $3 million to establish, but by 2009 the net annual value of its cadastral data to the state was estimated at $10 million (Freeman, 2011). The figures are illustrative rather than determinative, but they underline that structurally coherent land and property data can yield enduring public value across taxation, planning, environmental management and land-use policy.

Future consideration therefore centres not on “incremental improvement”, but on whether Wales should continue to rely on fragmented administrative data, or whether a more unified and authoritative cadastral infrastructure would better support reliable land-value estimation—and other policy functions—in the long run.

## CLARIFYING THE CONCEPTUAL PURPOSE AND DEFINITION OF “LAND VALUE”

A further consideration concerns the purpose for which land values are required. This project demonstrates that land value is not an observable quantity and cannot be recovered without modelling assumptions. The concept itself is contextual: the notion of “land value” relevant to insurance (the residual value absent structures), to

taxation (the social value of location), to planning (development potential), or to viability analysis (residual land value) can differ substantially.

Future work will therefore require clarity about:

- what definition of land value is sought,
- what separability assumptions are acceptable,
- what degree of complementarity between structure and location should be recognised, and
- how the chosen definition aligns with intended policy uses.

Without a clear conceptual purpose, methodological choices cannot be coherently evaluated.

## GOVERNING MODELLING DISCRETION AND METHODOLOGICAL CHOICES

All valuation approaches rely on judgement, whether made explicitly by modellers (as in hedonic regression), implicitly by algorithms (in machine-learning models), or embedded within formula design (as in Lot 3). This project highlights that no approach eliminates modelling discretion; each simply locates it differently.

If land-value estimation is to support operational decisions in future, Welsh Government may need to consider how modelling choices are documented, governed and updated, including assumptions about transformations, interactions, variable selection and calibration. This is particularly important because different modelling choices can yield different estimates even when based on the same underlying data.

## COMPLEMENTARITY BETWEEN MODELLING APPROACHES

Although the modelling approaches differ in structure and orientation, this project demonstrates that they can complement one another. Machine-learning models are well suited to detecting complex interactions and non-linearities, and may play a more substantial future role if policy definitions of land value require such effects to be captured. Hedonic regressions offer explicit control over functional form and decomposition. Formula-based models, while requiring calibration, can provide more easily communicated representations of spatial value patterns.

Future considerations therefore include how different approaches might be combined or sequenced depending on the conceptual definition of land value and the policy setting in which valuations are used.

## THE CONTINUING IMPORTANCE OF PUBLIC ENGAGEMENT

The behavioural evidence from Lot 5 shows that public acceptance of valuation approaches depends more on perceived intelligibility and fairness than on financial incentives or underlying technical performance. Any future work on land valuation in Wales would benefit from early and sustained engagement with citizens, not only to explain modelling choices but also to understand social perceptions of legitimacy.

A further consideration is the potential for the public to contribute directly to improving the underlying data. The Lot 5 dashboard already enabled participants to comment on the accuracy of parcel-level attributes, and many did so constructively. It would be a relatively modest extension to this platform to allow participants to propose corrections or supply missing information in a controlled way. Given the sheer volume of parcels in Wales and the local knowledge individuals hold about their own properties and surroundings, citizen-facilitated data validation could offer a valuable complement to official datasets, helping to identify mismatches, gaps or inconsistencies that arise from the fragmented data environment described above.

## REFERENCES


Ahlfeldt, G.M. & Wendland, N. (2009). 'Looming stations: Valuing transport innovations in historical context', *Economics Letters*, 105(1), pp. 97–99. doi:10.1016/j.econlet.2009.06.010.

ap Gwilym, R., Jones, E. & Rogers, H. (2020). *A technical assessment of the potential for a local land value tax in Wales.* Government Social Research Report 17/2020. Welsh Government, Cardiff. Available at: https://www.gov.wales/local-land-value-tax-technical-assessment

Bayerische Staatsregierung (no date). *Bayerisches Grundsteuergesetz (BayGrStG).* Available at: https://www.gesetze-bayern.de/Content/Document/BayGrStG/true (Accessed: 26 February 2026).

Biznes.gov.pl. *Podatek od nieruchomości – informacje dla przedsiębiorców.* Available at: https://www.biznes.gov.pl (Accessed: 26 February 2026).

Bundesministerium der Finanzen (no date). *Fragen und Antworten zur neuen Grundsteuer.* Available at: https://www.bundesfinanzministerium.de/Content/DE/FAQ/faq-die-neue-grundsteuer.html (Accessed: 26 February 2026).

Chambre des Députés (2023). *Projet de loi n° 8082 sur l'impôt foncier, l'impôt sur la mobilisation de terrains et l'impôt sur la non-occupation de logements.* Available at: https://wdocs-pub.chd.lu/ (Accessed: 26 February 2026).

Cummings, R.G., Holt, C.A. & Laury, S.K. (2009) 'Using economic experiments for policy making: An example from the Georgia irrigation reduction auction', *Journal of Economic Behavior & Organization*, 69(2), pp. 240–246. doi:10.1016/j.jebo.2008.02.010.

Danish Property Assessment Agency (Vurderingsstyrelsen) (no date). *Public property assessment and property tax – English guidance.* Available at: https://www.vurderingsportalen.dk/ejerbolig/english (Accessed: 26 February 2026).

Estonian Land Board (2025). *Landowner's Guide.* Geoportal of the Estonian Land Board. Available at: https://geoportaal.maaamet.ee/eng/spatial-data/cadastral-data/landowners-guide-p729.html (Accessed: 26 February 2026).

Finanzbehörde Hamburg (no date). *Grundsteuer – Informationen zum Hamburger Wohnlagenmodell.* Available at: https://www.hamburg.de/politik-und-verwaltung/behoerden/finanzbehoerde/themen/grundsteuer (Accessed: 26 February 2026).

Fonseca, M. & Grimshaw, S. (2017) ‘Do behavioural nudges work? A field experiment on rebates’, *Journal of Public Policy & Marketing*, 36(2), pp. 226–241. doi:10.1509/jppm.15.128.

Gibbons, S., Mourato, S. & Resende, G. (2014). ‘The amenity value of English nature: A hedonic price approach’, *Environmental and Resource Economics*, 57(2), pp. 175–196.

Hessisches Ministerium der Finanzen (no date). *Grundsteuer B in Hessen – Das Flächen-Faktor-Verfahren.* Available at: https://finanzamt.hessen.de/grundsteuer/grundsteuer-b-in-hessen (Accessed: 26 February 2026).

Jafary, P., Shojaei, D., Rajabifard, A. & Ngo, T. (2024). ‘Automated land valuation models: A comparative study of four machine learning and deep learning methods based on a comprehensive range of influential factors’, *Cities*, 151, 105115. doi:10.1016/j.cities.2024.105115.

Kuminoff, N.V., Parmeter, C.F. & Pope, J.C. (2010). ‘Which hedonic models can we trust? An assessment of the robustness of hedonic property value estimates’, *Journal of Environmental Economics and Management*, 60(3), pp. 145–160.

Landesamt für Steuern Niedersachsen (no date). *Grundsteuer B (Grundvermögen) – Das Flächen-Lage-Modell.* Available at: https://lstn.niedersachsen.de/steuer/grundsteuer/grundsteuer-b-grundvermogen-209755.html (Accessed: 26 February 2026).

Ma, J., Cheng, J.C.P., Jiang, F., Chen, W. & Zhang, J. (2020). ‘Analyzing driving factors of land values in urban scale based on big data and non-linear machine learning techniques’, Land Use Policy, 94, 104537. doi:10.1016/j.landusepol.2020.104537.

Malpezzi, S. (2002). ‘Hedonic pricing models: A selective and applied review’. In: O’Sullivan, T. & Gibb, K. (eds), *Housing Economics and Public Policy: Essays in Honour of Duncan Maclennan*. Blackwell Science, Oxford, pp. 67–89. doi:10.1002/9780470690680.ch5.

Montana Code Annotated (2025) §15-7-201. *Legislative intent – value of agricultural property.* Helena, MT: State of Montana. Available at: https://archive.legmt.gov/bills/mca/title_0150/chapter_0070/part_0020/section_0010/0150-0070-0020-0010.html?

Montana Code Annotated (2025) §15-6-133. *Class three property – description – taxable percentage.* Helena, MT: State of Montana. Available at:

https://archive.legmt.gov/bills/mca/title_0150/chapter_0060/part_0010/section_0330/0150-0060-0010-0330.html?

Montana State Library. *Montana Cadastral Framework Overview.* Available at: https://www.arcgis.com/home/item.html?id=f161a98b347b4cf29d371a6d7697912a (Accessed: 26 February 2026).

Montana Department of Revenue. *Property Valuation Process Presentation.* Outlining cost, sales comparison, and income approaches. Available at: https://archive.legmt.gov/content/Committees/Interim/2023-2024/Revenue/Meetings/September-2023/DOR-Property-Valuation-Process-Presentation.pdf (Accessed: 26 February 2026).

National Agency for Fiscal Administration (2015). *Fiscal Code (Law No. 227/2015), as amended: Title IX – Local Taxes.* Available at: https://static.anaf.ro/static/10/Anaf/Prezentare_R/Law227_11042018.pdf (Accessed: 26 February 2026).

National Tax Agency (NTA) (2025a). No. 4602 Tochi kaoku no hyōka [Valuation of land and buildings]. Available at: https://www.nta.go.jp/ (Accessed: 26 February 2026).

National Tax Agency (NTA) (2025b). Rosenka [Roadside land value map]. Available at: https://www.rosenka.nta.go.jp (Accessed: 26 February 2026).

Netherlands Enterprise Agency (RVO) (2022) 'Property tax (OZB)'. Available at: https://business.gov.nl/regulation/ozb/ (Accessed: 26 February 2026).

PwC (2019). 'Land Tax in Victoria'. Available at: https://www.pwc.com.au/tax/assets/stamp-duty/pdf/pwc-land-tax-in-victoria-june-2019.pdf (Accessed: 26 February 2026).

Quigley, J.M. (1985). 'Consumer choice of dwelling, neighbourhood and public services', *Regional Science and Urban Economics*, 15(1), pp. 41–63. doi:10.1016/0166-0462(85)90031-6.

Republic of Bulgaria (2006). *Local Taxes and Fees Act*, as amended. Available at: https://old.nra.bg/en/document?id=107 (Accessed: 26 February 2026).

Republic of South Africa (2004) Local Government: Municipal Property Rates Act 6 of 2004, as amended. Pretoria: Government Printer. https://www.gov.za/documents/local-government-municipal-property-rates-act-0? (Accessed: 26 February 2026).

Revenue NSW (2025). 'What is land tax?' and 'How land tax is calculated'. Available at: https://www.revenue.nsw.gov.au/ (Accessed: 26 February 2026).

Riigikogu (2022a). Land Valuation Act (consolidated text). State Gazette of Estonia, RT I, 10.03.2022, 2. Available at: https://www.riigiteataja.ee/en/eli/522032022004/consolide (Accessed: 26 February 2026).

Riigikogu (2022b). Land Tax Act (consolidated text). State Gazette of Estonia, RT I, 10.03.2022, 2. Available at: https://www.riigiteataja.ee/en/eli/530122024007/consolide (Accessed: 26 February 2026).

Rosen, S. (1974). 'Hedonic prices and implicit markets: Product differentiation in pure competition', *Journal of Political Economy*, 82(1), pp. 34–55. Available at: https://www.jstor.org/stable/1830899.

Sejm of the Republic of Poland. *Ustawa z dnia 12 stycznia 1991 r. o podatkach i opłatach lokalnych (Act of 12 January 1991 on Local Taxes and Fees).* Consolidated text available at: https://isap.sejm.gov.pl (Accessed: 26 February 2026).

Slovak Republic (2004). *Act No. 582/2004 Coll. on Local Taxes and Local Fee for Municipal Waste and Minor Construction Waste*, as amended. Available at: https://www.mfsr.sk/ (Accessed: 26 February 2026).

State Revenue Office Victoria (SRO Victoria) (2025) 'Site values and land tax'; Available at: https://www.sro.vic.gov.au/owning-property/land-tax/new-land-tax/site-values-and-land-tax? (Accessed: 26 February 2026).

StatsWales (2025). *Land Transaction Tax statistics, by transaction type and transaction description, latest period.* Welsh Revenue Authority. Available at: https://statswales.gov.wales/Catalogue/Taxes-devolved-to-Wales/Land-Transaction-Tax/landtransactiontaxstatistics-by-transactiontype-transactiondescription-latestperiod (Accessed: 26 February 2026).

Welsh Government (2021). *Future Wales: The National Plan 2040.* Cardiff: Welsh Government. Available at: https://www.gov.wales/future-wales-national-plan-2040.

Welsh Government (2024). *Planning Policy Wales: Edition 12.* Cardiff: Welsh Government. Available at: https://www.gov.wales/planning-policy-wales.

Xue, R., Gepp, A., O'Neill, T.J., Stern, S. & Vanstone, B.J. (2018). 'Financial literacy amongst elderly Australians', Accounting & Finance. doi:10.1111/acfi.12362.

Zhang, P., Hu, S., Li, W., Zhang, C., Yang, S. & Qu, S. (2021). 'Modeling fine-scale residential land price distribution: An experimental study using open data and machine learning', *Applied Geography*, 129, 102442. doi:10.1016/j.apgeog.2021.102442.

Zhou, X., Lennox, C., Li, Y. & Zeng, Y. (2025). ‘A systematic review of the urban land valuation literature’, *Journal of Accounting Literature*, 54, pp. 100–130. doi:10.1108/JAL-10-2024-0272.

# APPENDIX A: SUMMARY STATISTICS

**Table A1: Summary statistics for distance-to-amenities variables (meters)**

| Variable | Min | Max | Mean | Median | SD |
|---|---|---|---|---|---|
| distance_to_arts_centre | 7.49 | 56,040.26 | 17,433.27 | 13,625.34 | 14,630.88 |
| distance_to_attraction | 0.65 | 27,219.91 | 4,866.29 | 3,857.82 | 3,738.87 |
| distance_to_bus_stop | 1.00 | 9,259.77 | 170.15 | 119.62 | 273.48 |
| distance_to_castle | 3.21 | 37,611.98 | 5,820.83 | 4,906.56 | 4,553.76 |
| distance_to_college | 6.26 | 54,110.09 | 6,243.70 | 4,042.64 | 7,388.33 |
| distance_to_golf_course | 32.02 | 19,786.71 | 3,071.87 | 2,551.72 | 2,184.16 |
| distance_to_gp_surgery | 0.00 | 18,513.28 | 1,540.36 | 935.41 | 1,819.86 |
| distance_to_high_street | 0.84 | 19,743.45 | 1,366.72 | 837.20 | 1,697.03 |
| distance_to_hospital | 7.10 | 35,112.52 | 3,794.92 | 2,529.58 | 4,023.17 |
| distance_to_ice_rink | 184.10 | 236,220.45 | 74,776.19 | 42,124.10 | 71,631.09 |
| distance_to_in_town_retail | 3.39 | 38,839.01 | 7,108.65 | 5,331.78 | 6,731.91 |
| distance_to_library | 2.33 | 37,386.09 | 3,892.46 | 2,481.47 | 4,367.06 |
| distance_to_museum | 0.65 | 34,564.74 | 5,698.38 | 5,103.02 | 4,267.71 |
| distance_to_out_town_retail | 3.70 | 53,832.54 | 5,445.81 | 3,738.76 | 5,450.96 |
| distance_to_park | 8.27 | 23,409.26 | 1,016.21 | 474.29 | 1,794.02 |
| distance_to_pitch | 7.46 | 12,634.06 | 521.05 | 370.04 | 601.17 |
| distance_to_playground | 3.61 | 13,825.57 | 518.37 | 342.13 | 729.12 |
| distance_to_primary_school | 0.00 | 14,107.72 | 685.52 | 458.31 | 876.25 |
| distance_to_prison | 75.28 | 125,688.97 | 24,352.92 | 13,707.38 | 25,946.38 |
| distance_to_pub | 0.17 | 11,271.03 | 1,404.09 | 898.95 | 1,419.70 |
| distance_to_secondary_school | 32.39 | 33,987.50 | 2,429.67 | 1,452.63 | 2,948.95 |
| distance_to_sports_centre | 2.35 | 28,071.43 | 2,688.10 | 1,563.75 | 2,980.95 |
| distance_to_stadium | 38.66 | 66,154.43 | 8,633.04 | 5,402.93 | 10,661.62 |
| distance_to_swimming_pool | 7.42 | 23,762.42 | 3,480.68 | 2,362.26 | 3,359.61 |
| distance_to_theatre | 3.16 | 38,887.96 | 8,699.86 | 6,644.58 | 7,004.74 |
| distance_to_theme_park | 34.41 | 100,608.17 | 25,008.57 | 22,522.03 | 16,537.21 |
| distance_to_train_station | 4.12 | 36,195.50 | 3,289.31 | 1,708.00 | 4,567.81 |
| distance_to_university | 0.54 | 67,450.45 | 12,530.84 | 8,136.80 | 12,652.77 |
| distance_to_zoo | 16.25 | 75,081.65 | 17,866.06 | 17,031.46 | 10,082.71 |

**Table A2: Summary statistics for EPC numerical variables**

| Variable | Min | Max | Mean | Median | SD |
|---|---|---|---|---|---|
| CURRENT_ENERGY_EFFICIENCY | 1 | 148 | 63.63 | 66.0 | 14.57 |
| EXTENSION_COUNT | 0 | 4 | 0.53 | 0.0 | 0.76 |
| NUMBER_HABITABLE_ROOMS | 0 | 112 | 4.65 | 4.6 | 1.55 |
| NUMBER_HEATED_ROOMS | 0 | 90 | 4.00 | 4.0 | 2.15 |
| NUMBER_OPEN_FIREPLACES | 0 | 40 | 0.11 | 0.0 | 0.41 |
| POTENTIAL_ENERGY_EFFICIENCY | 1 | 191 | 79.42 | 81.0 | 10.33 |
| TOTAL_FLOOR_AREA | 0 | 133,244 | 145.46 | 92.0 | 253.65 |

**Table A3: Summary statistics for WIMD numerical variables**

| Variable | Min | Max | Mean | Median | SD |
|---|---|---|---|---|---|
| air_nitrogen_dioxide | 2.70 | 24.80 | 9.60 | 8.70 | 4.62 |
| air_pm10 | 6.40 | 15.10 | 10.57 | 10.60 | 1.69 |
| air_pm25 | 4.30 | 10.10 | 7.08 | 7.10 | 1.19 |
| broadband_unavailable_pct | 0.00 | 73.60 | 5.80 | 1.00 | 10.16 |
| cancer_incidence_rate | 217.80 | 987.90 | 611.26 | 605.10 | 84.96 |
| crime_asb_rate | 0.27 | 125.91 | 2.73 | 1.93 | 5.06 |
| crime_burglary_rate | 0.25 | 6.17 | 1.14 | 1.04 | 0.52 |
| crime_damage_rate | 0.16 | 13.06 | 1.18 | 0.95 | 0.91 |
| crime_theft_rate | 0.17 | 25.51 | 0.71 | 0.55 | 1.03 |
| crime_violent_rate | 0.24 | 76.84 | 2.71 | 2.14 | 3.42 |
| employment_deprivation_pct | 1.00 | 43.00 | 9.69 | 9.00 | 5.47 |
| fire_incidents_rate | 0.15 | 3.19 | 0.50 | 0.43 | 0.28 |
| flood_risk_score | 2.10 | 100.00 | 22.31 | 16.00 | 20.26 |
| foundation_phase_score | 83.00 | 114.00 | 104.62 | 105.00 | 3.57 |
| gp_chronic_condition_rate | 4.70 | 26.40 | 14.28 | 14.00 | 3.09 |
| gp_mental_health_rate | 5.40 | 46.40 | 22.91 | 22.50 | 5.03 |
| greenspace_access_pct | 0.00 | 100.00 | 77.43 | 84.80 | 23.36 |
| greenspace_score | -0.44 | 0.49 | 0.11 | 0.12 | 0.14 |
| hazardous_housing_pct | 2.10 | 55.70 | 17.41 | 15.40 | 8.59 |
| housing_disrepair_pct | 0.00 | 8.00 | 3.20 | 2.90 | 1.76 |
| income_deprivation_pct | 1.00 | 61.00 | 14.64 | 13.00 | 8.52 |
| ks2_score | 62.00 | 99.00 | 87.57 | 88.00 | 3.80 |
| ks4_score | 47.00 | 158.00 | 120.50 | 121.00 | 11.84 |
| ks4_to_he_pct | 2.60 | 70.50 | 31.65 | 30.70 | 11.61 |
| long_term_illness_rate | 9.50 | 40.70 | 22.33 | 21.60 | 5.24 |
| low_birth_weight_pct | 0.00 | 15.10 | 5.09 | 4.90 | 2.08 |
| no_quals_pct | 2.60 | 53.80 | 17.93 | 16.80 | 8.63 |
| overcrowded_households_pct | 0.00 | 25.63 | 5.22 | 4.57 | 3.22 |
| poor_housing_pct | 2.20 | 58.90 | 19.11 | 17.10 | 9.30 |
| premature_death_rate | 122.70 | 1,157.30 | 382.34 | 360.00 | 133.27 |
| repeat_absenteeism_pct | 0.20 | 21.50 | 4.82 | 4.30 | 2.76 |
| travel_private_foodshop_min | 1.00 | 20.00 | 3.58 | 3.00 | 2.49 |
| travel_private_gp_min | 1.00 | 34.00 | 6.57 | 5.00 | 4.30 |
| travel_private_library_min | 2.00 | 53.00 | 9.21 | 8.00 | 5.74 |
| travel_private_petrol_min | 2.00 | 41.00 | 7.42 | 7.00 | 4.08 |
| travel_private_pharmacy_min | 1.00 | 44.00 | 6.12 | 5.00 | 4.81 |
| travel_private_postoffice_min | 1.00 | 28.00 | 5.47 | 5.00 | 2.90 |
| travel_private_primary_min | 1.00 | 22.00 | 4.36 | 4.00 | 2.05 |
| travel_private_secondary_min | 2.00 | 47.00 | 10.96 | 9.00 | 6.38 |
| travel_private_sports_min | 1.00 | 45.00 | 9.20 | 8.00 | 5.93 |
| travel_public_foodshop_min | 13.00 | 180.00 | 29.21 | 22.00 | 22.03 |
| travel_public_gp_min | 14.00 | 177.00 | 36.79 | 29.00 | 25.51 |
| travel_public_library_min | 18.00 | 180.00 | 41.89 | 33.00 | 27.19 |
| travel_public_pharmacy_min | 15.00 | 180.00 | 34.74 | 27.00 | 25.91 |
| travel_public_postoffice_min | 15.00 | 180.00 | 33.33 | 27.00 | 21.62 |
| travel_public_primary_min | 16.00 | 180.00 | 31.53 | 26.00 | 19.23 |
| travel_public_secondary_min | 20.00 | 180.00 | 51.32 | 41.00 | 29.84 |
| travel_public_sports_min | 18.00 | 180.00 | 51.67 | 37.00 | 34.94 |

**Table A4: Summary statistics for the 9 identified LSOAs – distance-to-amenities variables**

| Variable | Min | Max | Mean | Median | SD |
|---|---|---|---|---|---|
| distance_to_arts_centre | 356.75 | 31,655.50 | 12,968.13 | 12,950.67 | 11,245.62 |
| distance_to_attraction | 55.33 | 27,219.91 | 5,131.72 | 500.29 | 7,508.53 |
| distance_to_bus_stop | 5.63 | 6,363.24 | 151.06 | 104.84 | 252.15 |
| distance_to_castle | 69.10 | 16,996.41 | 2,725.23 | 712.74 | 3,560.01 |
| distance_to_college | 297.28 | 49,549.32 | 10,591.09 | 8,827.77 | 11,943.42 |
| distance_to_golf_course | 79.73 | 17,276.70 | 4,706.90 | 4,408.56 | 3,419.06 |
| distance_to_gp_surgery | 20.18 | 12,514.60 | 1,462.10 | 810.43 | 1,940.20 |
| distance_to_high_street | 3.35 | 12,519.44 | 1,147.43 | 272.69 | 2,084.55 |
| distance_to_hospital | 30.60 | 30,503.10 | 6,711.47 | 1,579.88 | 7,599.95 |
| distance_to_ice_rink | 2,865.28 | 193,225.00 | 66,899.99 | 35,931.31 | 68,706.66 |
| distance_to_in_town_retail | 6.60 | 30,866.09 | 8,911.78 | 8,903.47 | 8,627.22 |
| distance_to_library | 418.71 | 29,681.45 | 7,010.40 | 3,429.09 | 7,874.58 |
| distance_to_museum | 93.70 | 30,745.98 | 7,762.92 | 7,041.67 | 7,654.17 |
| distance_to_out_town_retail | 230.76 | 36,797.32 | 7,483.05 | 2,186.39 | 9,368.50 |
| distance_to_park | 23.38 | 9,383.20 | 908.30 | 401.90 | 1,419.72 |
| distance_to_pitch | 17.98 | 6,411.30 | 612.62 | 434.86 | 742.88 |
| distance_to_playground | 17.06 | 5,926.55 | 667.46 | 410.65 | 844.15 |
| distance_to_primary_school | 21.62 | 6,556.46 | 710.95 | 546.27 | 730.60 |
| distance_to_prison | 186.09 | 89,513.44 | 24,154.47 | 10,814.17 | 31,452.23 |
| distance_to_pub | 1.51 | 10,189.52 | 1,256.80 | 506.77 | 1,544.17 |
| distance_to_secondary_school | 627.45 | 13,553.94 | 4,068.56 | 2,559.28 | 2,934.87 |
| distance_to_sports_centre | 40.66 | 9,204.63 | 4,153.34 | 4,648.92 | 3,087.27 |
| distance_to_stadium | 89.87 | 64,788.34 | 14,642.54 | 6,707.42 | 21,105.55 |
| distance_to_swimming_pool | 40.07 | 13,054.09 | 4,734.01 | 4,355.83 | 3,870.46 |
| distance_to_theatre | 17.17 | 30,569.67 | 9,904.49 | 8,319.88 | 9,366.22 |
| distance_to_theme_park | 8,729.22 | 99,268.90 | 32,754.27 | 24,226.72 | 27,229.71 |
| distance_to_train_station | 43.04 | 12,163.25 | 3,012.02 | 1,786.48 | 3,003.55 |
| distance_to_university | 0.54 | 67,450.45 | 14,592.34 | 7,032.89 | 19,297.10 |
| distance_to_zoo | 2,674.01 | 75,081.65 | 25,020.58 | 20,202.33 | 17,052.00 |

**Table A5: Summary statistics for the 9 identified LSOAs – EPC numerical variables**

| Variable | Min | Max | Mean | Median | SD |
|---|---|---|---|---|---|
| CURRENT_ENERGY_EFFICIENCY | 1 | 92 | 65.26 | 69 | 15.88 |
| EXTENSION_COUNT | 0 | 4 | 0.39 | 0 | 0.72 |
| NUMBER_HABITABLE_ROOMS | 1 | 20 | 4.16 | 4 | 1.76 |
| NUMBER_HEATED_ROOMS | 0 | 20 | 3.45 | 3 | 2.23 |
| NUMBER_OPEN_FIREPLACES | 0 | 8 | 0.09 | 0 | 0.38 |
| POTENTIAL_ENERGY_EFFICIENCY | 1 | 119 | 78.46 | 81 | 10.71 |
| TOTAL_FLOOR_AREA | 0 | 10,913 | 214.59 | 87 | 480.74 |

**Table A6: Summary statistics for the 9 identified LSOAs – WIMD numerical variables**

| Variable | Min | Max | Mean | Median | SD |
|---|---|---|---|---|---|
| air_nitrogen_dioxide | 3.10 | 24.60 | 12.96 | 7.40 | 8.89 |
| air_pm10 | 6.70 | 14.60 | 11.58 | 11.10 | 2.45 |
| air_pm25 | 4.60 | 9.40 | 7.52 | 7.10 | 1.52 |
| broadband_unavailable_pct | 0.00 | 45.70 | 20.59 | 22.70 | 18.51 |
| cancer_incidence_rate | 367.20 | 695.90 | 550.76 | 574.30 | 115.87 |
| crime_asb_rate | 0.54 | 125.91 | 33.16 | 4.31 | 47.47 |
| crime_burglary_rate | 0.63 | 2.93 | 1.58 | 1.72 | 0.77 |
| crime_damage_rate | 0.32 | 13.06 | 4.06 | 1.62 | 4.59 |
| crime_theft_rate | 0.41 | 25.51 | 6.89 | 0.84 | 9.59 |
| crime_violent_rate | 0.70 | 76.84 | 21.16 | 6.04 | 28.52 |
| employment_deprivation_pct | 1.00 | 17.00 | 5.91 | 5.00 | 4.92 |
| fire_incidents_rate | 0.22 | 3.01 | 0.99 | 0.53 | 1.02 |
| flood_risk_score | 7.60 | 89.20 | 32.81 | 31.00 | 22.07 |
| foundation_phase_score | 83.00 | 111.00 | 102.37 | 105.00 | 9.86 |
| gp_chronic_condition_rate | 4.70 | 17.40 | 11.85 | 12.40 | 3.77 |
| gp_mental_health_rate | 5.40 | 28.80 | 16.20 | 15.80 | 6.88 |
| greenspace_access_pct | 24.90 | 100.00 | 75.73 | 76.10 | 22.32 |
| greenspace_score | -0.18 | 0.46 | 0.06 | 0.07 | 0.20 |
| hazardous_housing_pct | 9.50 | 35.30 | 17.45 | 16.40 | 6.91 |
| housing_disrepair_pct | 2.30 | 6.60 | 3.99 | 3.70 | 1.16 |
| income_deprivation_pct | 1.00 | 20.00 | 7.86 | 8.00 | 5.98 |
| ks2_score | 62.00 | 93.00 | 80.83 | 86.00 | 11.88 |
| ks4_score | 104.00 | 137.00 | 120.70 | 121.00 | 9.42 |
| ks4_to_he_pct | 12.90 | 40.00 | 27.31 | 27.30 | 7.88 |
| long_term_illness_rate | 12.30 | 28.10 | 17.79 | 18.50 | 4.64 |
| low_birth_weight_pct | 1.70 | 7.40 | 4.97 | 5.00 | 1.80 |
| no_quals_pct | 2.60 | 24.10 | 10.71 | 7.70 | 7.28 |
| overcrowded_households_pct | 1.20 | 13.75 | 7.44 | 6.69 | 3.77 |
| poor_housing_pct | 10.00 | 37.90 | 19.46 | 19.20 | 7.10 |
| premature_death_rate | 230.20 | 488.50 | 322.77 | 309.90 | 64.64 |
| repeat_absenteeism_pct | 0.80 | 6.60 | 4.25 | 4.90 | 1.99 |
| travel_private_foodshop_min | 1.00 | 12.00 | 3.24 | 2.00 | 2.53 |
| travel_private_gp_min | 4.00 | 13.00 | 6.48 | 6.00 | 2.45 |
| travel_private_library_min | 3.00 | 31.00 | 7.30 | 5.00 | 6.54 |
| travel_private_petrol_min | 3.00 | 12.00 | 6.99 | 8.00 | 2.37 |
| travel_private_pharmacy_min | 3.00 | 14.00 | 5.12 | 4.00 | 2.93 |
| travel_private_postoffice_min | 3.00 | 13.00 | 5.47 | 5.00 | 2.62 |
| travel_private_primary_min | 2.00 | 12.00 | 4.48 | 4.00 | 2.16 |
| travel_private_secondary_min | 8.00 | 32.00 | 17.83 | 16.00 | 7.47 |
| travel_private_sports_min | 3.00 | 32.00 | 11.26 | 9.00 | 8.99 |
| travel_public_foodshop_min | 18.00 | 120.00 | 28.45 | 20.00 | 22.06 |
| travel_public_gp_min | 23.00 | 127.00 | 34.15 | 27.00 | 22.08 |
| travel_public_library_min | 18.00 | 123.00 | 36.43 | 28.00 | 26.20 |
| travel_public_pharmacy_min | 20.00 | 125.00 | 31.19 | 23.00 | 22.70 |
| travel_public_postoffice_min | 21.00 | 123.00 | 32.88 | 25.00 | 21.70 |
| travel_public_primary_min | 20.00 | 99.00 | 29.44 | 27.00 | 16.77 |
| travel_public_secondary_min | 34.00 | 144.00 | 65.55 | 50.00 | 36.48 |
| travel_public_sports_min | 22.00 | 144.00 | 52.00 | 38.00 | 33.71 |

## APPENDIX B: LOT 1 TECHNICAL ASSESSMENT

The preferred model is estimated using Ordinary Least Squares within a hedonic framework, with the natural logarithm of total transaction price specified as the dependent variable. The estimation includes structural, locational, environmental, and socio-economic controls, with parcel area incorporated explicitly on the right-hand side. The logarithmic specification was selected following comparative testing of alternative outcome variables and functional forms.

Transaction-level land data exhibit pronounced right skew and scale-dependent variance. Residual variance typically increases with transaction magnitude, resulting in heteroskedasticity and inefficient estimation. The logarithmic transformation reduces skewness, compresses the influence of extreme high-value observations, and stabilises variance across the fitted range. It also permits proportional interpretation of coefficients. Comparative testing of a linear level specification confirmed greater residual dispersion and sensitivity to extreme transactions relative to the logarithmic form.

The model achieves an $R^2$ of 0.441 and an adjusted $R^2$ of 0.440. The negligible difference between these values indicates limited inflation of explanatory power from model dimensionality. In the context of parcel-level hedonic modelling, where prices are influenced by numerous unobserved factors, this degree of explanatory performance is consistent with empirical expectations. Variation in $R^2$ across property types and local authorities reflects differences in market structure and dispersion rather than specification instability.

Predictive accuracy was assessed in level space by back-transforming fitted values into total transaction price. The overall Root Mean Squared Error is £315,552. Given the scale-dependence of absolute error, performance was examined across transaction value deciles. Median relative error ranges between approximately 14 and 18 per cent in the middle deciles, increasing to around 34 per cent in the top decile and 74 per cent in the bottom decile. The increase in error at the upper end of the distribution is gradual rather than disproportionate, indicating that the logarithmic specification maintains broadly proportionate performance.

Residual diagnostics were conducted using a random subsample of 300,000 observations. The mean log residual is 0.0005 and the median is approximately zero, indicating no systematic over- or under-prediction. The 5th and 95th percentiles are −0.69 and 0.59 respectively, and the correlation between fitted values and residuals is effectively zero at −0.001. These statistics suggest that residual variance does not systematically increase across the fitted range, consistent with substantial mitigation of heteroskedasticity under the chosen functional form.

Stability testing using a 200,000 observation subsample indicates that approximately 89 per cent of coefficients retain the same sign as in the full-sample model, demonstrating robustness of estimated relationships. The logarithmic total price specification therefore represents a balanced and statistically coherent modelling choice.

# APPENDIX C: LOT 5 ANALYSIS

The residents that completed our experiment made decisions over three alternatives: Hedonic Pricing, Machine Learning and Equation Based LVMs. These three alternatives do not have a natural ordering. Thus, our dependent variable, the resident's choice of LVM, is categorical. As such, we use multinomial logit regressions to analyse the data parametrically.

The coefficient estimates produced by multinomial logits are not intuitive, unlike the coefficient estimates from OLS regressions. For simplicity, we report the *relative risk ratios*. The coefficients can be interpreted in the following, simple, way:

- A coefficient that is statistically significant and greater than 1 implies the outcome is more likely, relative to the baseline.
- A coefficient that is statistically significant and less than 1 implies the outcome is less likely, relative to the baseline.

In all the estimates we present, we use the *Equation based* estimate as the baseline from which we compare the other outcomes.

We estimate five regressions. In the first regression, we only include the impact of the *High* treatment on the dependent variable, relative to the baseline. To outline how the coefficients are interpreted, we will discuss this regression.

The left most column outlines the outcome of the dependent variable – Hedonic and Machine Learning. The second column then outlines the effect of the variables on that outcome, relative to the baseline. As can be seen, the *High Treatment* variable has a positive effect on the outcome *Hedonic*, but this estimate is less than 1. It is also not statistically significant at the 5% level – hence, it is estimated *not* to impact the likelihood of a resident selecting Hedonic, relative to the baseline. The same is true for Machine Learning – the coefficient estimate on *High Treatment* is positive, but less than 1 (0.6) and not statistically significant.

The term *Constant* outlines the overall likelihood of the outcome relative to the baseline. As can be seen in regression (1), this is positive for Hedonic, large than 1 (3.822) and significant at the 5% level – hence, the Hedonic pricing method is more likely to be chosen than the baseline (Equation based modelling). In each of the four subsequent regressions, we add additional variables to examine how they correlate with the dependent variable. Regression (2) adds accuracy, Regression (3) adds Comprehension, Regression (4) adds the region variables and Regression (5) adds type of resident. As can be seen across all regressions, none of the included variables (Accuracy, Comprehension, the Welsh region variables) are statistically significant. Hence, the estimates suggest that none of the included variables have any impact on the choice of LVM.

**Table C1: Multinomial Logit Estimates**

| LVM | | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|---|
| Hedonic | | | | | | |
| | Constant | 3.822** | 7.28** | 8.65* | 8.99* | 12.63 |
| | | (2.34) | (6.94) | (10.99) | (11.73) | (22.45) |
| | *High Treatment* | 0.668 | 0.678 | 0.678 | 0.66 | 0.653 |
| | | (3.822) | (0.257) | (0.257) | (0.255) | (0.258) |
| | *Accuracy* | | 0.919 | 0.921 | 0.93 | 0.933 |
| | | | (0.086) | (0.086) | (0.088) | (0.089) |
| | *Comprehension* | | | 0.948 | 0.934 | 0.906 |
| | | | | (0.246) | (0.244) | (0.239) |
| | *Mid Wales* | | | | 1.53 | 1.6 |
| | | | | | (0.802) | (0.856) |
| | *North Wales* | | | | 1.59 | 1.77 |
| | | | | | (1.17) | (1.318) |
| | *Swansea Bay* | | | | 0.608 | 0.584 |
| | | | | | (0.278) | (0.269) |
| | *Home Owner* | | | | | 0.648 |
| | | | | | | (0.787) |
| | *Tenant* | | | | | 0.908 |
| | | | | | | (1.128) |
| | *Prefer not to say* | | | | | 1.32 |
| | | | | | | (1.83) |
| Machine Learning | | | | | | |
| | Constant | 3.7** | 4.65 | 8.698* | 8.778* | 10.78 |
| | | (2.33) | (4.62) | (11.34) | (11.742) | (19.89) |
| | *High Treatment* | 0.6 | 0.6 | 0.606 | 0.57 | 0.533 |
| | | (2.35) | (0.236) | (0.238) | (0.227) | (0.218) |
| | *Accuracy* | | 0.971 | 0.976 | 0.979 | 0.981 |
| | | | (0.096) | (0.096) | (0.098) | (0.098) |
| | *Comprehension* | | | 0.82 | 0.791 | 0.772 |
| | | | | (0.216) | (0.209) | (0.206) |
| | *Mid Wales* | | | | 1.55 | 1.588 |
| | | | | | (0.863) | (0.896) |
| | *North Wales* | | | | 2.48 | 2.66 |
| | | | | | (1.861) | (2.01) |
| | *Swansea Bay* | | | | 1.02 | 0.962 |
| | | | | | (0.472) | (0.451) |
| | *Home Owner* | | | | | 0.89 |
| | | | | | | (1.135) |
| | *Tenant* | | | | | 0.925 |
| | | | | | | (1.21) |
| | *Prefer not to say* | | | | | 2.013 |
| | | | | | | (2.89) |

Note: Equation based modelling taken as the baseline. Standard errors in brackets. *, **, *** denotes statistical significance at the 10%, 5% and 1% level. Low treatment taken as the baseline. 201 observations in all regressions.