Model specification document

Introduction

This document sets out the data utilised in development of the Automated Valuation Model for a future revaluation of Council Tax in Wales, as well as the core spatial regression model specification and an overview of the comparables model.

Data

Data Utilised in Development

The variables used in the AVM model include property attributes, locations and sales details. While much of this data is sourced from VOA records, the VOA supplements this with data available across Government, including from the Office for National Statistics, HM Land Registry, and through the Public Sector Geospatial Agreement from Ordnance Survey.

A full table of data and sources is given below:

Category	Data	Detail	Source
Sales	Sale Price & Date	The date a property sold and the price it sold for	VOA sales data; HM Land Registry Price Paid Data
	Sale Price Adjustment	The House Price Index gives the relative change in value from one time period to another, enabling sales to be adjusted to the valuation date	Office for National Statistics / HM Land Registry House Price Index
Location	Property Coordinates	The latitude and longitude of a property.	Ordnance Survey AddressBase Premium; VOA records; National Statistics Postcode Lookup (NSPL)
	Billing Authority	The Billing Authority that a property is located in.	VOA Records
Property Attributes	Dwelling Type	For example, a flat, an inner-terrace or a detached house.	VOA Records
	Dwelling Group	This broadly describes the architectural style of the property.	
	Dwelling Age	The time period when the property was built.	
	Dwelling Area	The size of the property in m^2. We measure properties in line with our code of measuring practice – guidance on how we measure different properties is available on GOV.UK.	
	Parking Facilities	The type of parking at a property. For example, a garage or open parking space.	
	Subsidised Housing Indicator	This indicates if a property has been purpose-built as social housing by a Local Authority,	

		Housing Association or other public body.	
	Bedroom Count	We count all bedrooms, even if they are not currently being used as bedrooms. This includes box rooms and bedrooms being used as studies.	
	Bathroom Count	We only list a bathroom if it has at least 3 fittings, for example, toilet, basin and shower. This means a toilet with just a basin would not be included.	
	Plot size	The area of the curtilage of the property in m^2.	Land Registry Title Polygons and Ordnance Survey National Geographic Database

Additionally, the VOA explored the use of Energy Performance Certificate data and building footprints from Ordnance Survey National Geographic Database, however neither of these were included in the final model.

Updating VOA Records

An exercise to enhance VOA's property data was undertaken to update our records in line with the latest property attributes. VOA records were only updated through manual desk-based review. This activity was undertaken in the same way that we maintain the existing Council Tax Lists. A range of sources are utilised by VOA staff. These include:

- overhead and street view imagery from search engine maps as well as aerial photography from Ordnance Survey available to the VOA through the Public Sector Geospatial Agreement
- publicly available plans, plans obtained with the consent of the owner/taxpayer, and during agreed inspections of properties
- sale listing particulars, available online or through third party services

Sales Adjustment

The purpose of the model is to predict the value of a property at the Antecedent Valuation Date (AVD) which for purpose of developing our model was set at 01 April 2023. However, the model needs sufficient data to be trained to provide accurate valuations. It therefore needs to include sales over a reasonable time period. Changes in prices vary over time and location so this needs to be accounted for. The ONS publish the House Price Index (HPI), broken down by Billing Authority and (high level) Dwelling Types. We adopted these rather than deriving our own index.

By experimenting with different years of data and understanding the impact on the model we decided to use data from three-years prior to the AVD for houses and six-years for flats (where there are fewer properties and therefore fewer sales).

Model

The AVM is a series models incorporating a core regression model and a comparables model. The regression model is further split into houses and flats.

The following table indicates the property data used in each model:

	Regression Model – Houses	Regression Model – Flats	Comparables model
Dwelling Type	Yes	Yes	Yes
Dwelling Group	No	No	Yes
Dwelling Age	Yes	Yes	Yes
Dwelling Area	Yes	Yes	Yes
Parking Facilities	Yes	Yes	No
Subsidised Housing Indicator	Yes	Yes	Yes
Bedroom Count	No	No	Yes
Bathroom Count	No	Yes	No
Plot size	Yes	No	No

Regression Model Specifications

The regression models use multiple regression, combined with a Gaussian Markov Random Field for the spatial elements.

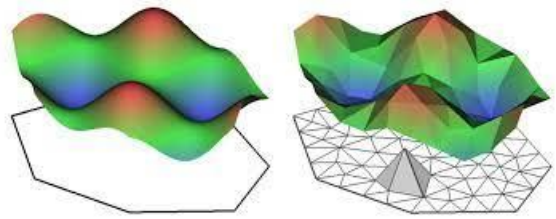
The model specifications are given in the following table. A few things should be noted:

- Variables prefaced with "log" have had the logarithm taken. This is a mathematical transformation that can make the data more suitable for modelling and therefore yield better results. This also results in a multiplicative model where the coefficients give impacts in terms of percentage changes.
- Variables suffixed with "std" have been standardised, which expresses the variable as how many standard deviations it is away from the mean. It greatly simplifies the computation of some models and so is required.
- Most categorical variables are grouped.
- A "x" indicates an interaction term, whereby coefficients are given for the combination of variables, for instance Dwelling Area x Dwelling Type means there is a different coefficient for Dwelling Area for each Dwelling Type.

	Houses	Flats
Fixed Effects	log_Dwelling Area_std x Dwelling Type + Age x Dwelling Type + Subsidised Housing Indicator + Parking Facilities + log_Plot size_std + Billing Authority x Dwelling Type	log_Dwelling Area_std * Dwelling Type + Age + Subsidised Housing Indicator + Parking Facilities + Bathroom Count

This model uses the coordinates of sales properties to estimate a continuous field of location adjustments (see left side of Figure 1). The modelling approach used (Gaussian Markov Random Field model) is frequently used in spatial modelling. It enables the property values to be modelled as if they vary continuously over all of Wales rather than just varying between fixed geographical regions. In practice, the modelling approach approximates the smooth, continuous variation in property values using a mesh of triangles across Wales (see right side of Figure 1).

Figure 1: Illustration of the continuous field and its approximation with triangular mesh (Figure taken from https://ourcodingclub.github.io/)



Comparables Model

The comparables model scores how similar two properties are, accounting for their property attributes and location. Individual attributes are scored, with a weighted sum combining these scores to give an overall comparability score. Additionally, the score weights the usefulness of a sale by considering its proximity to the AVD.

Weights were determined based on valuer professional judgement around how important each element is in choosing comparable properties. The following table gives the weights for each element and the method used for scoring:

Attribute	Weighting	How scored
Characteristics	**57%**	-
Social Housing	23%	Categorical variable
Type	10%	Categorical variable
Area	8%	Broadly triangular distribution, but with longer tail
Bedrooms	8%	Categorical variable
Age	4%	Categorical variable
Group	4%	Categorical variable
Distance	**33%**	Broadly S curve distribution reaching 0 at distance to 100th closest comparable
Closeness of Sale to AVD	**10%**	Triangular distribution with AVD as centre and ±3 years for houses and ±6 years for flats

The score for each element falls within the range 0 to 1. For categorical variables, a cross-tabulation showing the scores for each possible pair of values for the attribute is used; for instance, a detached house is fully comparable to another detached house (getting a score of 1) whilst a semi-detached house is much less comparable to a detached house (getting a score of 0.5). Continuous variables are scored using a formula which will assign a score between 0 and 1 depending on how similar the two values for the variable are.