



## Summer data challenge entry by Benjamin L. Moore<sup>1</sup>

### OVERVIEW

I analysed the **London house prices** dataset in several innovative ways. Firstly *fractal context* shows how house sales within an arbitrary area relate to those of neighbouring districts, up and down the postcode hierarchy. Secondly, I fit *ARIMA models* to forecast an area's price trend into the future, with an eye to suggesting profitable investments. Finally, these and other metrics were combined into an *investment grade* for a given postcode area. Combined this set of analyses forms the basis of **datarea**, a product aimed at "democratising real estate investment", offering individuals and businesses quantitative real estate insights for data-driven investment decisions.

A full report with additional visualisations is online at [blm.io/datarea](https://blm.io/datarea) and scripts to reproduce all analyses are available from [github.com/blmoore/summerdatachallenge](https://github.com/blmoore/summerdatachallenge).

### ANALYSES

When considering a property in a given area, you may ask a realtor questions like: how do prices here compare to surrounding postcodes? How have prices increased over time? And how have prices changed over time? A real estate professional can give an opinion based on their own limited sample size, but with a large dataset we can address these questions quantitatively and visualise the results.

#### Fractal context

Letting and sales sites may currently list some recent sales, but it's not currently possible to see a quantitative overview of property prices in a region, within the context of its sector, district and postcode.

These questions can be answered with a clear and intuitive data visualisation of price distributions in the postcode hierarchy — we call this the *fractal context* of a property price (Figure 1).

A specific postcode or area can be marked within its distribution, giving a clear view of a properties relative pricing which cannot be gleaned just by browsing recent sales. The example view (Figure 1) combines ranked violin density plots and a geo-heatmap for a clear, intuitive view of a locations price context.

#### ARIMA modelling

Given a time series of housing prices, it's of interest to investors to speculate on potential investment returns for a given property area. This

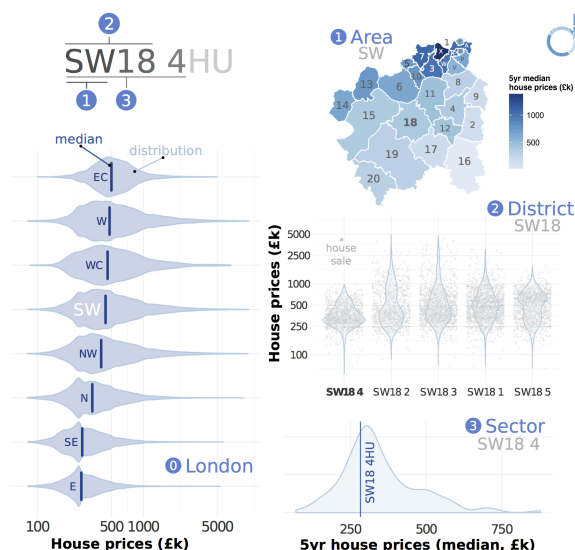


Figure 1: Fractal context of postcode SW18 4HU in Wandsworth, South London.

<sup>1</sup> MRC Human Genetics Unit, University of Edinburgh, Scotland, United Kingdom

can be done through time series analysis using the autoregressive integrated moving average (ARIMA) model (Figure 2). The model fitting procedure included regularisation that allowed a drift term to model non-stationary trends, as well as coefficients capturing periodicity or seasonal effects. Potential investment returns could be optimised by selecting areas with maximal growth projections.

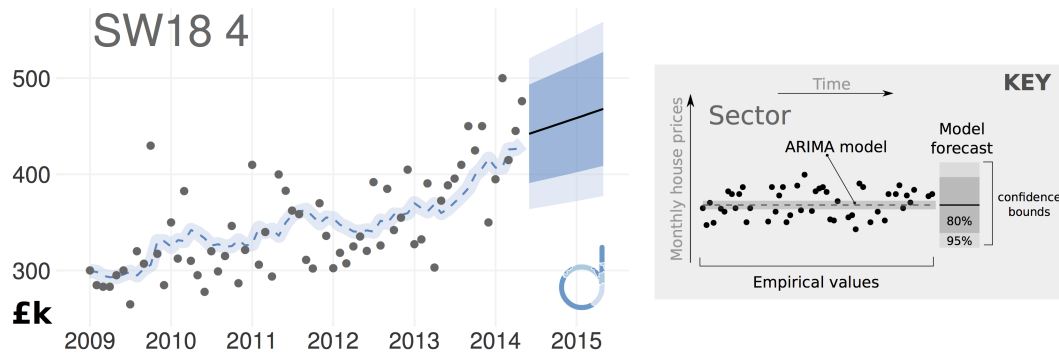


Figure 2: Example ARIMA forecast of median house prices for a specific postcode.

#### Investment grading

Combining growth forecasts and other derived metrics from this dataset allow an approximate investment grading, relative to other postcodes within the area covered by this dataset. This intuitive output metric can help prioritise areas ripe for property investment.

As an example, Figure 3 shows the top 5 best-ranked postcode sectors in which to invest, according to their 12-month price growth forecast and historical annualised volatility — thus combining the empirical data with theoretical model outputs. These include South London suburbs of Upper Tooting and Morden, Walthamstow, as well as Brockley, whose train station within SE4 1 was linked with the London overground network in 2010.

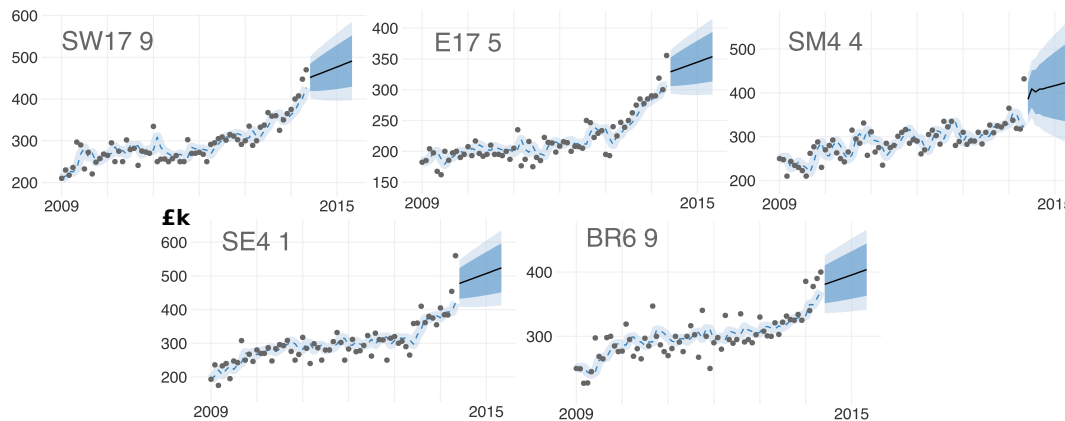


Figure 3: Top 5 property investment sectors by projected returns and historical volatility.

The rankings used in Figure 3 made use of the drift parameter of ARIMA models, combined with statistical volatility over the five year house prices dataset. Both parameters were converted to quantiles and weighted equally, with those sectors with the largest projected growth and lowest historical volatility receive the highest investment grade.\*

\* Data intended to assist investors and does not constitute investment advice; independent advice should be sought where appropriate.

## VALUE GENERATION

### Socio-economic value

#### Initial product

dotarea would first be released as both a web service and associated apps for mobiles and tablets. These would offer limited analysis and visualisation for free, with user analytics and feedback helping to validate the business model.

A small data science team will gradually expand the range of analyses offered, integrating novel public and privately-acquired datasets and release them under a rolling subscription model, aimed at private landlords and property speculators, as well as professional real estate investment trusts (REITs). Significant revenue could be generated through tailored partnerships with funds and high net-worth individuals.

Additionally the business could pivot to an analytics provider for existing letting and sales sites, offering unique data insights indirectly to the consumer market, or as independent fund consultants for the financial services. Having produced a first-to-market innovative product positions dotarea to succeed as a provider of advanced real estate investment analytics.

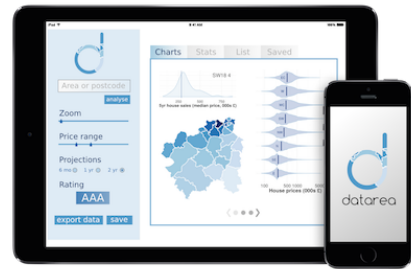


Figure 4: iOS app mock-up.



### Business development

An enhanced overview of this entry is available online at: [blm.io/dotarea](https://blm.io/dotarea)