Summer data challenge

Benjamin L. Moore

20th October 2014

OVERVIEW

I analysed the **London house prices** dataset is several innovative ways. Firstly *fractal context* shows how house sales within an arbitrary area relate to those if neighbouring districs, 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. Together this set of analyses forms the basis of datarea, a startup aimed at *democratising real estate investment*, offering individuals and businesses quantitative property area insights for data-driven investment decisions.

A full report with additional visualisations is online at **blm.io/datarea** and scripts to reproduce all analyses are available from **github.com/blmoore/summerdatachallenge**.

A map of the house prices dataset (brightness is proportional to price and density).

ANALYSES

When considering a property in a given area, you may ask a realtor questions like: how do prices here compare to surrounding postcodes, and how have they 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.

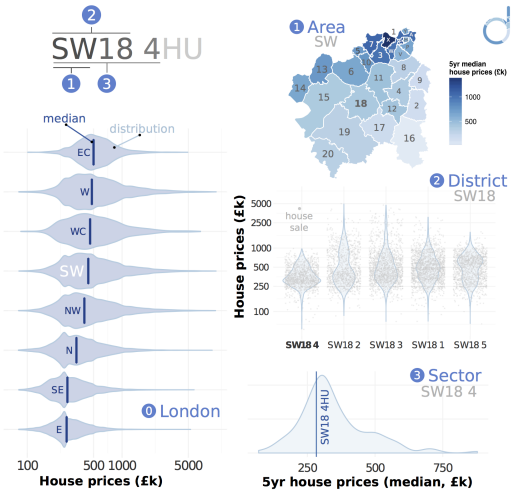

Figure 1: Fractal context of postcode SW18 4HU in Wandsworth, South London.

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 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

[1] *MRC Human Genetics Unit, University of Edinburgh, Scotland, United Kingdom*

trends, as well as coefficients capturing periodicity or seasonal effects. Potential investment returns could be optimised by selecting areas with maximal growth projections.

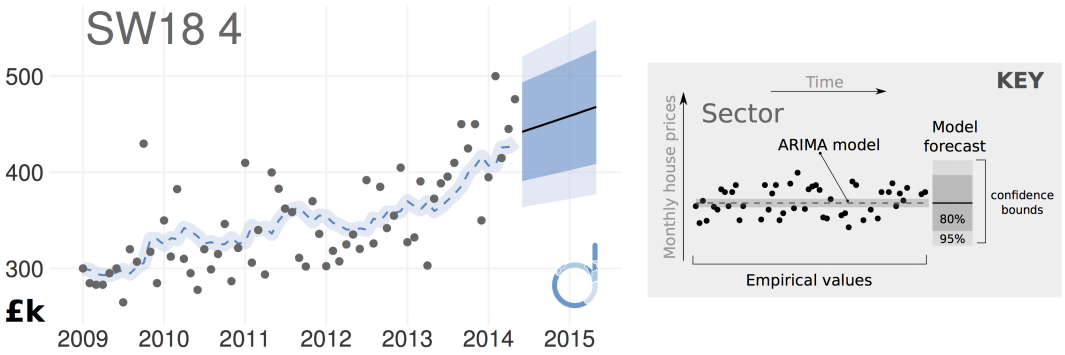

Figure 2: Example ARIMA forecast of median house prices for a specific postcode.

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, as well as Walthamstow in East London, and Brockley, whose train station within SE4 1 was linked with the London Overground network in 2010.

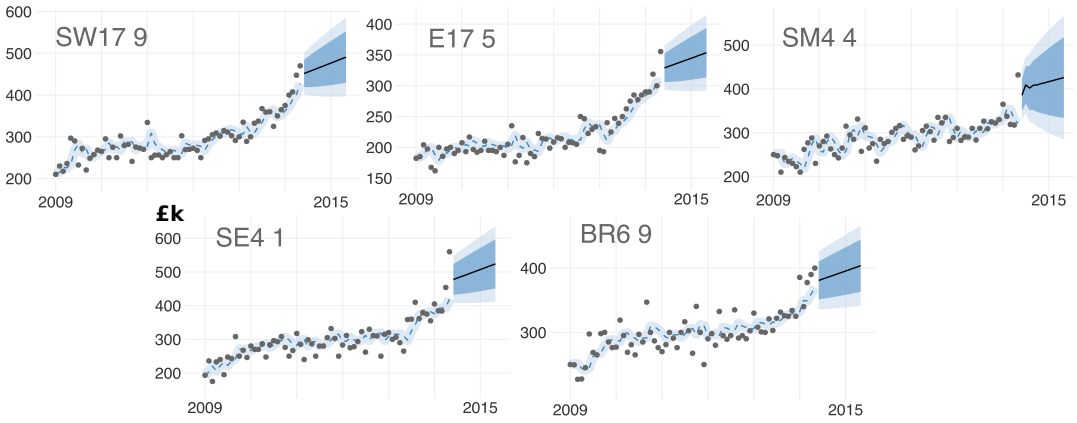

Figure 3: Top 5 property investment sectors by projected returns and historical volatility.

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; those sectors with the highest projected growth and least historical volatility receive the highest investment grade (on a scale of AAA, AA, A, BBB . . . C).*

Returning to our original example, we can conclude that SW18 4HU is a promising choice of area for property investment. It's approximately median-priced for the sector, which is the cheapest of SW18 (in turn part of the desirable SW London area; Figure 1). It has good growth projections of XX in the next 12 months (Figure 2), and when combined with volatility and compared with all $\approx$ 2600 sectors in the dataset, receives the top AAA investment grade, placing it in the 90^{th} percentile for predicted price growth and low historical volatility.

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

VALUE GENERATION

The above-described analyses form the basis of datarea, a start-up aimed at "democratising real estate investment" by making deep-dive analytics data available both to a consumer and professional investment market. Initially datarea is a provider of information aimed at investors but cannot dispense investment advice until authorised to do so by the FCA. Three initial areas of interest to commercialise and generate social and economic value are listed:

1. Analytics provider for online house sales and lettings agents

datarea can provide **real-time property area analytics** to existing market leaders in **online sales and lettings**.

Potential partners.

There are currently no mainsteam property sales or rental sites that attempts to provide any quantitative insights to contextualise a property's list price. Some list recent nearby sales in plain text, but the wealth of historical sales data available allows for much richer displays (as demonstrated with the fractal context view, *above*), which offer valuable and practical insights to prospective homebuyers.

This added feature would help differentiate a given sales and letting site as well as increase the brand positioning of datarea within the real estate sector. Contextual maps could be drawn as interactive D3.js visualisations and neatly integrated with a typical property search. Additionally, web analytics data from our partnered site could be fed-back to help further enhancement and expansion of the data-driven insights and visualisations we offer.

2. Direct-to-consumer demo application

datarea would release both a **web service** and associated **mobiles and tablet applications**. These would offer limited analysis and visualisation without any fee, with collected user data and feedback helping to validate the business model and develop our software.

Releasing these products aligns with our aims of democratising real estate investment. Amateur property investors and prospective landlords can make data-driven decisions, the likes of which are likely already employed by large investment trusts and funds.

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.

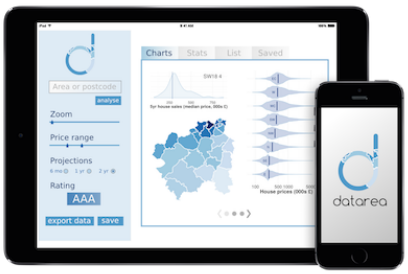

Figure 4: iOS app mock-up.

3. Real estate data science consultancy

Having established a property area analytics brand, datarea will look to develop business relationships with real estate investment funds and high net-worth individuals as a **quantitative property investment consultancy** and independent analyst for the financial services. Our domain expertise and existing partnerships and products would place us as a market leader in this undeveloped segment.

An enhanced online version of this short report is available at: blm.io/datarea