Bay area blues: the effect of the housing crisis

Hadley Wickham, David Poole and Deborah F. Swayne

January 7, 2009

Introduction

There has been much talk of the housing crisis in the media, and much speculation who has been worst hit and how long it might last. In this chapter we're going to take a more quantitative look at the housing crisis by exploring data about the sales of half a million homes in the Bay Area. We are going to use a few simple statistical tools, but mostly we will focus on graphical displays of the data.

To perform our investigation, we will mainly use R, a statistical programming and data analysis environment. We'll discuss R and a few other we used in more depth later on, but for now lets dive into the data. We'll start by discussing how we got the data, the cleaning and transformation we performed before we could start any analysis, and then we'll show some of the interesting things that we uncovered.

How did we get the data?

Finding relevant datasets for a particular problem is challenging and requires a lot of exploration and investigation. We were particularly luckily to stumble over weekly house sales updates for the Bay Area produced by the SF Chronicle, `http://www.sfgate.com/homesales/`. Initially we had planned on scraping the data off the website, but a little detective work revealed that the data is already available in a machine readable format. Each human readable (html web page) weekly summary is built from a machine readable text file that looks like this:

```
rowid: 1
county: Alameda County
city: Alameda
newcity: 1
zip: 94501
street: 1220 Broadway
price: $509,000
br: 4
lsqft: 4420
bsqft: 1834
year: 1910
```

Each week's worth of data is available at a url of the form `http://www.sfgate.com/c/a/year/month/day/REHS.tbl`. This is pretty convenient, and only requires generating a list of all Sundays from the first on record, 2003/04/27 (which we found on the archive page), to the most recent, 2008/11/16 (at the time of analysis). With list of dates in hand, we then generated a list of urls in the correct format and then downloaded them with the command line tool `wget`. We used wget because it can easily resume where it left off if interrupted: this saves a lot of time when you're moving from place to place on a laptop.

With all the data on a local computer, the next step was to convert the data into a standard format. We use csv (comma separated values) for most datasets: although there's no standards document that exactly describes the structure of csv files, for most statistical datasets it's not complicated and every statistical package (and Excel!) can read it in without problems. This gives us a file like this:

```
county,city,zip,street,price,br,lsqft,bsqft,year,date,datesold
Alameda County,Alameda,94501,1220 Broadway,509000,4,4420,1834,1910,2003-04-27,NA
Alameda County,Alameda,94501,429 Fair Haven Road,504000,4,6300,1411,1964,2003-04
-27,NA
Alameda County,Alameda,94501,2804 Fernside Boulevard,526000,2,4000,1272,1941,200
3-04-27,NA
Alameda County,Alameda,94501,1316 Grove Street,637000,3,2700,1168,1910,2003-04-2
7,NA
```

This is a little less human readable (there's much less white space), but it's a standard format that we can easily work with. Another minor advantage is that's much smaller than the original: 45 vs 90 megabytes. If you look closely at the sample data you might notice something that needs some explanation: the **NA**s. NA stands for not applicable, and is the sentinel value that R uses to represent missing values. Missing values have special semantics which, by default, will propagate missingness throughout a summary: + NA = NA, 5 ¡ NA = NA and so on. We always need to make a deliberate decision to drop the missing values from an analysis.

It takes just a few minutes to parse the files for all 293 weeks and get `house-sales.csv`, a csv file with 521,726 observations and 11 variables. It took much more time to tweak the parser to get all the edge cases right: we needed to convert prices to regular numbers (by removing $ and ,), parse the dates into a consistent format, and fill in missing values for fields that didn't occur in all of the tables. This is common in data analysis: the time taken to compute the answer is totally overwhelmed by the time necessary to develop the correct approach.

Geocoding

When we first looked at this data, we thought it would be really important to geocode all 436,107 unique addresses. That is, we wanted to associate a latitude and longitude with each address so that it would be easy to explore fine-grained spatial effects. In the end, we didn't end up using this extra data as much as we thought we would, but it's still an interesting challenge: how can you geocode nearly half a million addresses?

We started by looking at the well-known web services provided by google and yahoo. These were no good for two reasons: strict daily limits on the number of requests and heavy licensing restrictions on the resulting data. The daily limits meant that it would take well over a month to geocode all the addresses, and then the licensing would mean that we couldn't publish our results! After a lot of googling, we found a fantastic open service, the USC WebGIS Services, provided by the GIS research laboratory at the University of Southern California (Goldberg and Wilson, 2008). This service is free for non-commercial use and makes no restrictions on the uses to which you can put the data. It has no daily usage cap, but there is an implicit cap caused by the speed: we could only geocode about 80,000 addresses per day, so it took us around 5 days to do all 400,000. The disadvantage of this free service is that the quality of the geocoding isn't quite as good (they only use publicly available address data), but the creators were very helpful and have published an excellent free introduction to the topic in Goldberg (2008). (We suspect that the software included a bug during the period we used it, so that large numbers of addresses were falsely assigned to the Los Angeles area and elsewhere around the state, and we remapped these addresses using another free on-line service at gpsvisualizer.com.)

As well as latitude and longitude, the USC results also include an indication of how accurate the geocoding is. 10% percent of the addresses were located exactly based on property boundaries (extremely accurate), another 75% percent were located by interpolating between the numbers at each end of the block (very accurate), 7% to the centre of the zip code (not very accurate) and the remainder were only located to the centre of the city or not at all.

Data checking

It's worth spending a lot of time with this data to ensure it's accurate. If it's not, any problems will propagate through to the rest of our analysis. We discovered quite a few unusual locations! Errors in geocoding come from a number of sources: There are typographical errors in the addresses, and new buildings are often not in public databases. The addresses in San Jose were a particular challenge. Sales are listed for several "towns" that are not acknowledged in any mapping sites we could find, so we assume they are informal names for neighbourhoods: North, South, East and West San Jose, Berryessa, Cambrian, and a few others. Our debugging process included drawing simple maps by county and town to identify the addresses that had been located far afield.

Where possible we tried to correct any errors. When that wasn't possible, we used R's missing values to indicate that we don't really the latitude and longitude. This is a better approach than throwing out bad matches, because we need varying levels of accuracy for different purposes: city level accuracy is fine when we are comparing cities, will want address level when we are looking within a city, or focusing on purely geographical comparisons. Using missing values for lat and long ensures that any location with a suspicious geocoding will be dropped from analyses that uses latitude and longitude, but included in all others.

Analysis

When starting an analysis, it's best to start with a very broad overview. Given that we're interested in the housing crisis, we'll start by looking at the weekly number of sales and average price. Once we have a feel for the overall patterns, we'll start breaking the data up into smaller pieces and seeing how they compare to each other and to the overall patterns. We will look at three breakdowns, by house price (from most expensive to least), and spatially, both between cities and within a single city (San Francisco).

Figure 1 shows weekly sale numbers and average prices for the 293 weeks of the data. There are a lot of interesting patterns in this data. The effect of the housing crisis on average prices is striking, with an increasing trend until June 2007 and then a sharp drop until the present day — the housing crisis. Sales show a different pattern. From mid 2006, we see a gradual decrease in sales volume, and then an increasing trend in early 2008. Maybe by this point house prices had dropped enough that people were shopping for bargains again.

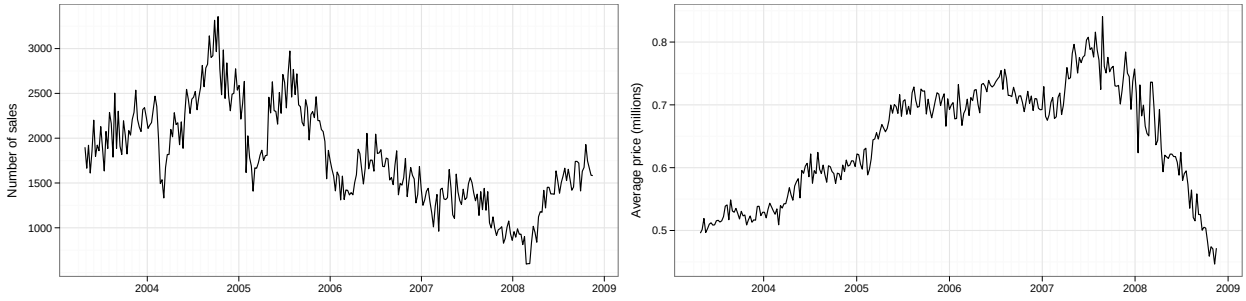

Figure 1: Weekly sales (left) and average prices (right).

Adjusting for inflation

The data is over a relatively short period of time (almost 6 years), but we might wonder if it's necessary to adjust for inflation, to ensure that the prices paid in 2003 are comparable to the prices paid today. A commonly used reference for calculating inflation is the consumer price index (CPI) produced by the Bureau of Labor Statistics, `http://www.bls.gov/CPI`. The CPI calculates the price of a weighted "basket" of frequently purchased consumer goods. This price is calculated monthly, and we will use the West coast series, series CUUR0400SA0. We use this series to adjust for inflation as follow. We want to adjust all values to 2003 dollars, so we divide each CPI value by its value in March 03. This operation is also known as indexing. This series gives the relative worth of a 2003 dollar at each point in time and makes it easy to read the effect of inflation from the graph: a value of 1.1 represents a cumulative inflation of 10% from the start of the data. Indexing is a very useful technique and we'll use it throughout our analysis. Figure 2 shows the CPI-based inflation measurement and the effect of adjusting prices for inflation. Inflation has been steadily climbing over the last five years, and failing to adjust for inflation makes the increasing trend prior to mid 2007 look more pronounced. However, inflation adjustment is complicated because housing prices form a major part of the CPI, and because of this we chose not to inflation adjust the prices.

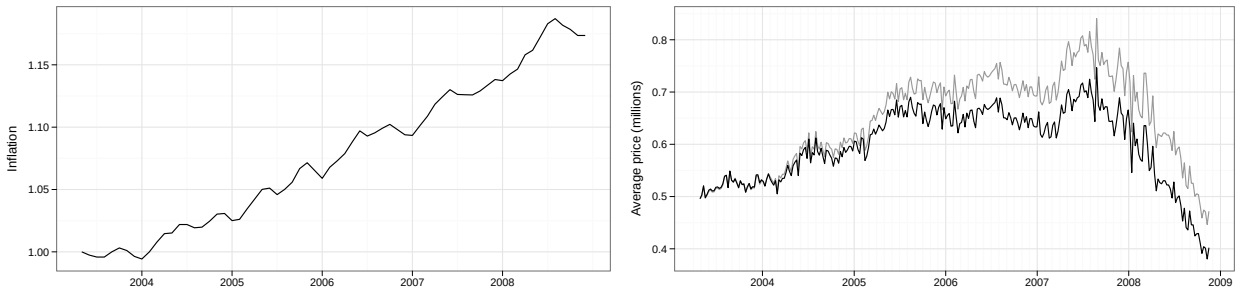

Figure 2: (Left) Inflation, indexed at 1 at start of series. (Right) Inflation adjusted prices in 2003 dollars (black), and unadjusted prices (grey). Failing to adjusting for inflation makes the rise look steeper, but has little effect on the decline.

With this basic overview in hand, it's time to start drilling into the details. In the following sections we break the house sales in to smaller groups, first by price and then by location. We'll see whether the housing crisis has affected all equally, or some more than others.

The rich get richer and the poor get poorer

Has the housing crisis equally affected the rich and the poor? Has the effect of the crisis been to improve or worsen the relative equality of these two groups? In this section, we will explore how the crisis has affected the distribution of housing prices. A big caveat is that we're looking at the Bay Area, so homes will be more expensive than many other places in the country, but we still might expect to see some relative inequalities.

To start our exploration, we calculate price deciles for each month. The deciles are the nine prices that 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90% of houses cost less than. This is a succinct summary of the *distribution* of the prices for each month: instead of just looking at the average price, we are looking at nine numbers that summarise the complete distribution of the prices.

Figure 3 shows how these deciles have changed over time. The top line is the ninth decile, the price that 90% of houses are less than, and the bottom line is the first decile, the price that only 10% of houses are cheaper than. The line in the middle is the median, the price which divides the houses into halves, half cheaper and half more expensive. The lines are coloured from dark to light from expensive to cheap. Each line follows a similar pattern, and we can see the effect of the housing bubble in mid 2007, particularly in

the most expensive houses.

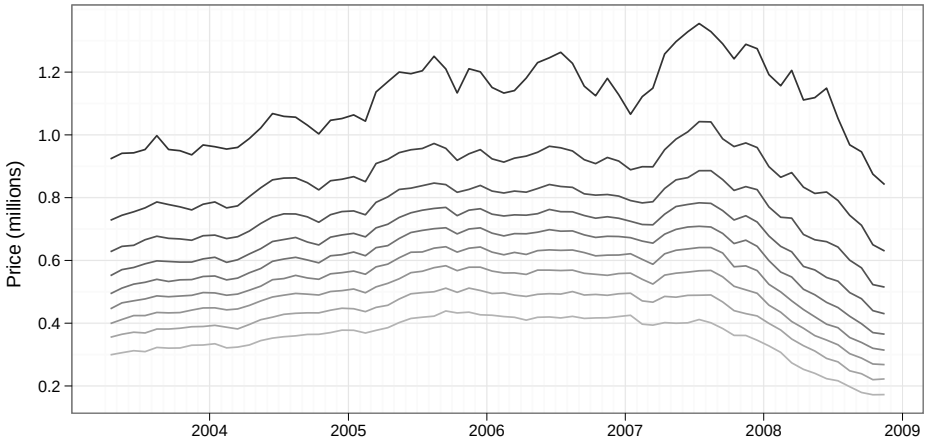

Figure 3: Monthly average house price within each decile. Lower lower deciles have lighter colours.

This plot lets us compare the absolute values of each decile, but maybe it is more appropriate to look at the relative prices: how have the prices changed proportionately? One way to look at the relative price is to compare each decile to its initial value. To do this we index each decile, dividing each series by its initial price, just as we did for the CPI. Figure 4 shows these indices. Each decile starts at one, and we can see the relative change in prices over time. What's interesting in this plot is that the cheaper houses (the lighter coloured lines) seem to peak higher and earlier (mid 2005), and then drop more rapidly. The cheapest houses lost 43% of their 2003 value compared to only 9% for the most expensive houses.

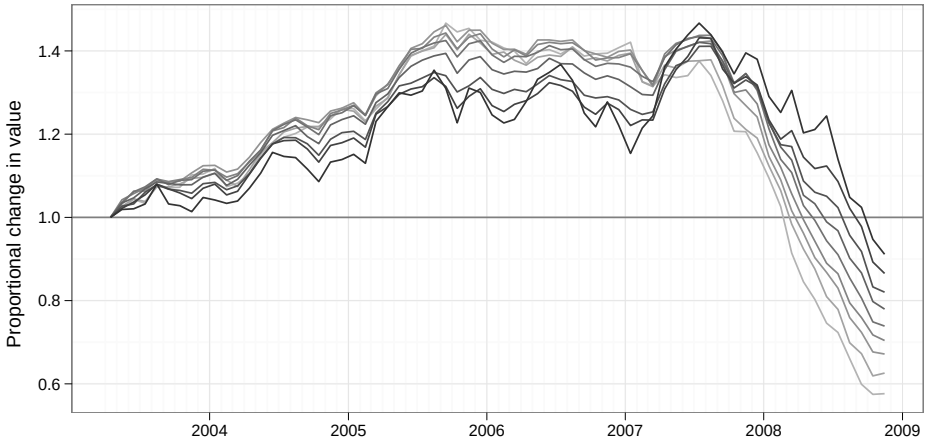

Figure 4: Indexed house price within each decile. The average price of cheaper houses peaked higher and earlier, and fell more steeply.

Another way to look at this inequality is Figure 5. Here we have divided all the prices by the price of an average (median) home. The values now represent a proportion of the median house price: a price of 1.2 represents a price 20% higher than the median, and 0.8 20% lower. Since the beginning of 2007 (before the housing crisis) the relative inequality has benn growing. This is interesting, and would be interesting to

follow up on — has it preceded other crises?

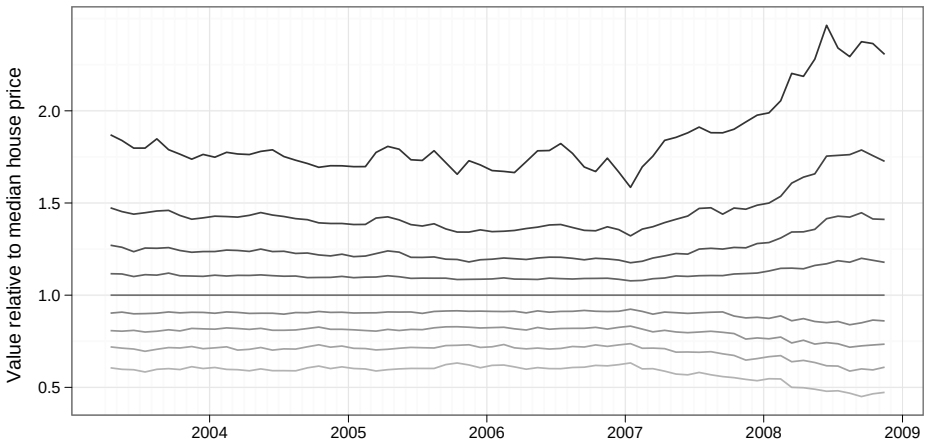

Figure 5: House prices, relative to the price of the median priced home. The disparity in home prices is increasing.

Geographic differences

We've just seen how the housing crisis has affected expense and cheap houses differently, but how does has it affected different cities? In this section we'll explore how the effect of the housing crisis has varied over different cities in the bay area. Because we are looking at the average prices, we'll need to pick out cities with a decent amount of data. We decided to focus on all cities with more than an average of 10 sales per week. This gives us 58 cities (out of an 245, 24%) with 428,415 sales (out of 521,726, 82%).

We then calculated the average weekly house price. Figure 6 shows these prices, with each city drawn with a different line. Statisticians have an evocative name for this type of plot: the spaghetti plot. It's very hard to see anything in the big jumble of lines. One way to improve this is to smooth each line, to focus on the long-term trend and remove the short-term variation that we're not so interested in.

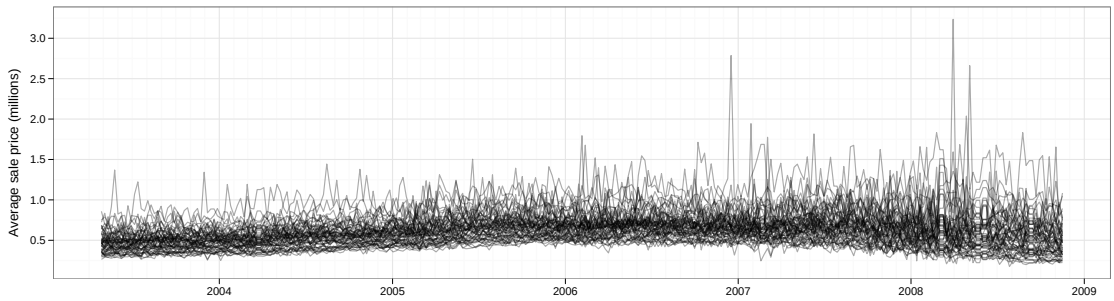

Figure 6: Average sale price for each week for each city. This type of plot is often called a spaghetti plot.

There are many different ways to create smooth curves, and for our purposes of them most of them are fine. We used generalised additive models (GAM), a generalisation of linear models (Wood, 2006), that fit smooth curves by optimising the trade off between being close to the data and being very smooth. The effect is to remove noisy short-term effects and focus on the long-term trend. This is what we want: we're not interested in daily or weekly changes, just the long-term changes related to the housing crisis.

The left part of Figure 7 shows the the result of this smoothing. This is a big improvement and we can now actually see some patterns! Note the big difference in scales between this plot and the first: smoothing the data has removed the large spikes which represent the sales of few very expensive houses. We'll also index each city like we indexed each decile: dividing out the starting price puts each city onto an common scale and allows us to focus on the changes. This is show on the right of Figure 7.

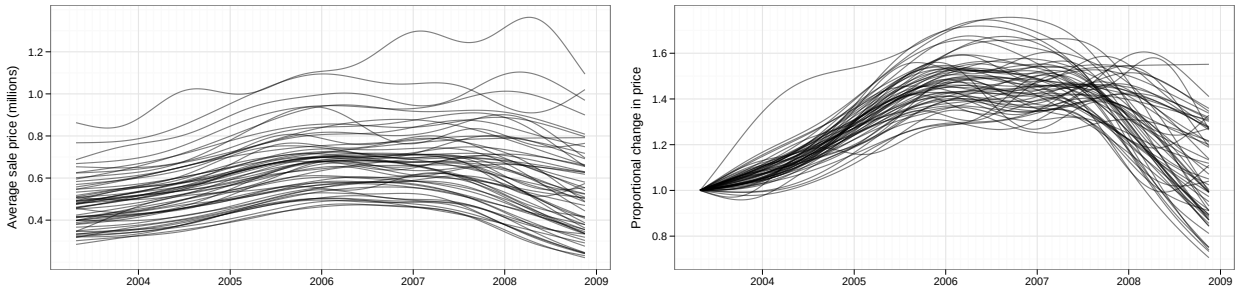

Figure 7: Smoothed city-level weekly average sale prices. Compared to the non-smoothed version it's easier to see the long-term trends, but it's still not particularly easy.

There is a still a lot of variation, but we can start to see a pattern of increasing values until mid 2007, and then decreasing values afterwards. To get any further, we need to look at the cities individually, as in Figure 11. This takes up a lot of space, but if you have a big screen or a good printer it's really worthwhile. We can pick out some interesting patterns: Berkeley and San Francisco show less of a peak and less of a drop, and Mountain View seems not all affected by the crisis. Other cities, like Oakley, Vallejo, and San Pablo, show big peaks and big drops.

After a few false starts we realised that there was one main feature that seemed to distinguish the different cities: the difference between prices at the peak of the boom and the depth of their most recent plummet. We create a new variable drop, that is the difference between the price in Feburary 2006 (at the height of the boom) and the price in November 2008 (the current doldrums). Figure 8 groups the cities by this new variable. The divisions are arbitrary, but you can see how the cities in each group follow a similar pattern: the bigger the boom, the bigger collapse. This suggests that this single number does a good job of summarising the different effects of the housing crisis.

We have determined that some cities showed a different pattern to others, but why? Plotting the geographic pattern, as in Figure 9 doesn't reveal anything particularly striking except that the worst hit towns tend to be the north and the east. This doesn't offer much in the way of explanatory power, so we looked for other variables that might help us to understand what's going on.

The census quickfacts site, e.g. `http://quickfacts.census.gov/qfd/states/06/0649670.html`, provides a number of interesting demographic variables for each city. Unfortunately the data isn't available in an easily downloadable format, so we had to resort to another set of scripts to scrape the data and convert it into csv. There were also some differences in what constituted a city in the census data and so we could only match 46 out of the full 58 cities. The cities that we couldn't match either too small, or included in a larger city in the census data.

Looking at the demographic variables revealed that the most affected cities have a high percentage of babies and children, bigger households, fewer bachelors degrees, and longer commutes. Most significantly these cities also have lower average incomes, which is probably the factor that drives all of the other relationships. Figure 10 shows three scatterplots showing the relationship between the drop and income, percent of college graduates and commute time. The relationship with commute time is weak, but striking in that all of the cities with particularly long commute times (¿25 minutes) have particularly large drops. It appears that

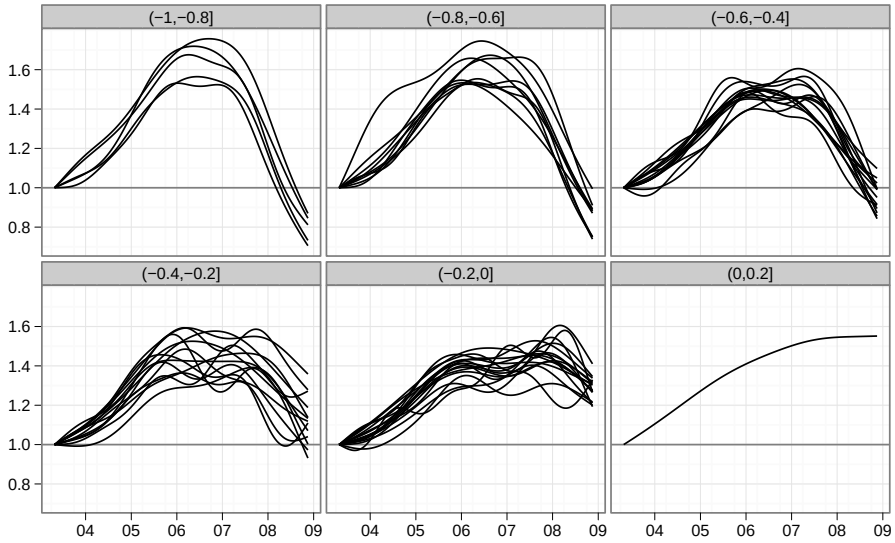

Figure 8: A separate plot for each 0.2 interval drop. The patterns within each group are similar, suggesting that this one number does a good job of

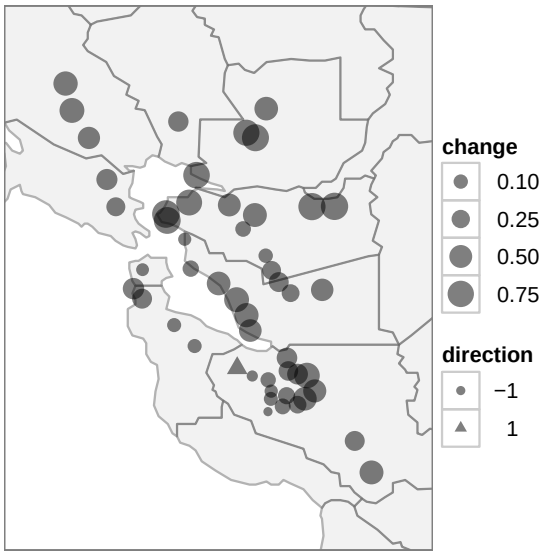

Figure 9: The geographic distribution of price drops.

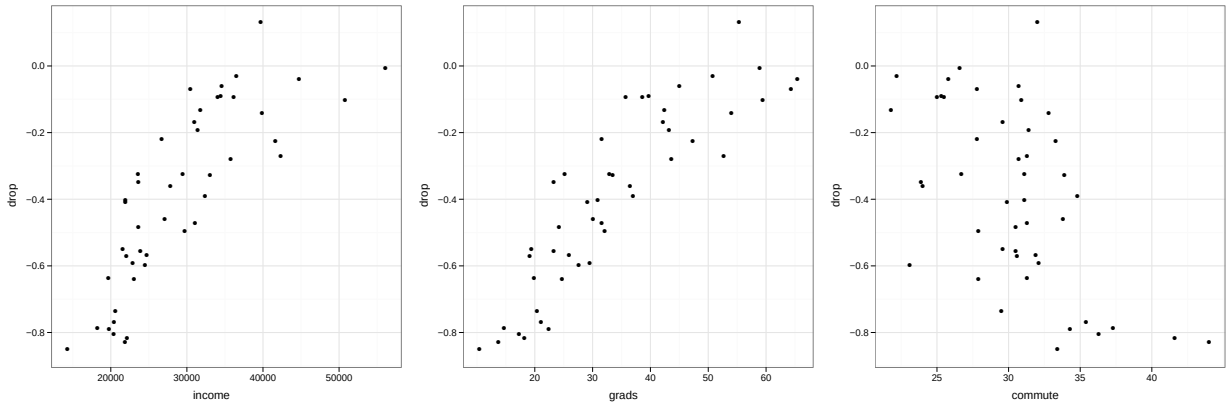

Figure 10: From left to right, the relationship between the drop in house prices and average income, percent of college graduates and average commute time.

the housing crisis has been relatively more damaging to poorer areas.

Exploring San Francisco

Having explored the difference between cities, we turned to look at a single city in more detail. We chose San Francisco because it's the city that we were most familiar with, and it has some iconic features that it should be easy for others to identify as well. We started our exploration by extracting all addresses within San Francisco that were geocoded to block interpolation or better, giving us a total of 25,377 addresses. With this data we did a very simple plot: a scatterplot of the latitudes and longitudes, Figure 12. For the residential parts of SF, this gives an amazingly detailed picture. We can the orientation of the streets, the waterfront boundaries and parks. Our view of some parts, like downtown, is more patchy because there are few residential homes there.

One problem with this plot is we lose all feel for how many sales are in each location. Figure 13 shows two attempts to recapture the information. On the left, we have a bubbleplot with the size of the location proportional to the number of sales. We now get quite a different view of the downtown: there are a lot of sales there. Looking closer at the data reveals that these are apartment buildings with hundreds of apartments. On the right, we have divided SF into squares of 0.005 latitude and longitude and counted the number of houses in each bin. This gives us higher level view showing where the majority of houses are.

Using that same binning, we calculate the mean and coefficient of variation of the house prices. The coefficient of variation is the standard deviation divided by the mean. We use it here because a variation of $100,000 is relatively much more important when houses are cheap compared to when they are expensive. Figure 14 shows the geographic distribution of these two summary statistics. We can see the most expensive houses border the Presidio and coast to the North of SF. There also seems to be a peak in the Southwest - this is the affluent St Francis wood area, near San Francisco State University. There is an interesting geographic trend in the coefficient of variation: it seems to increase towards the Northwest. We don't know enough about San Francisco to guess as to why this occurs.

Conclusions

We have used relatively simple statistical tools like indexing, quantiles, smoothing and binning to explore a large complex data set. We started with broad summaries and then dig down to explore the details. This is a large dataset and we have only just scratched the surface. If the data has caught your interest and you'd like

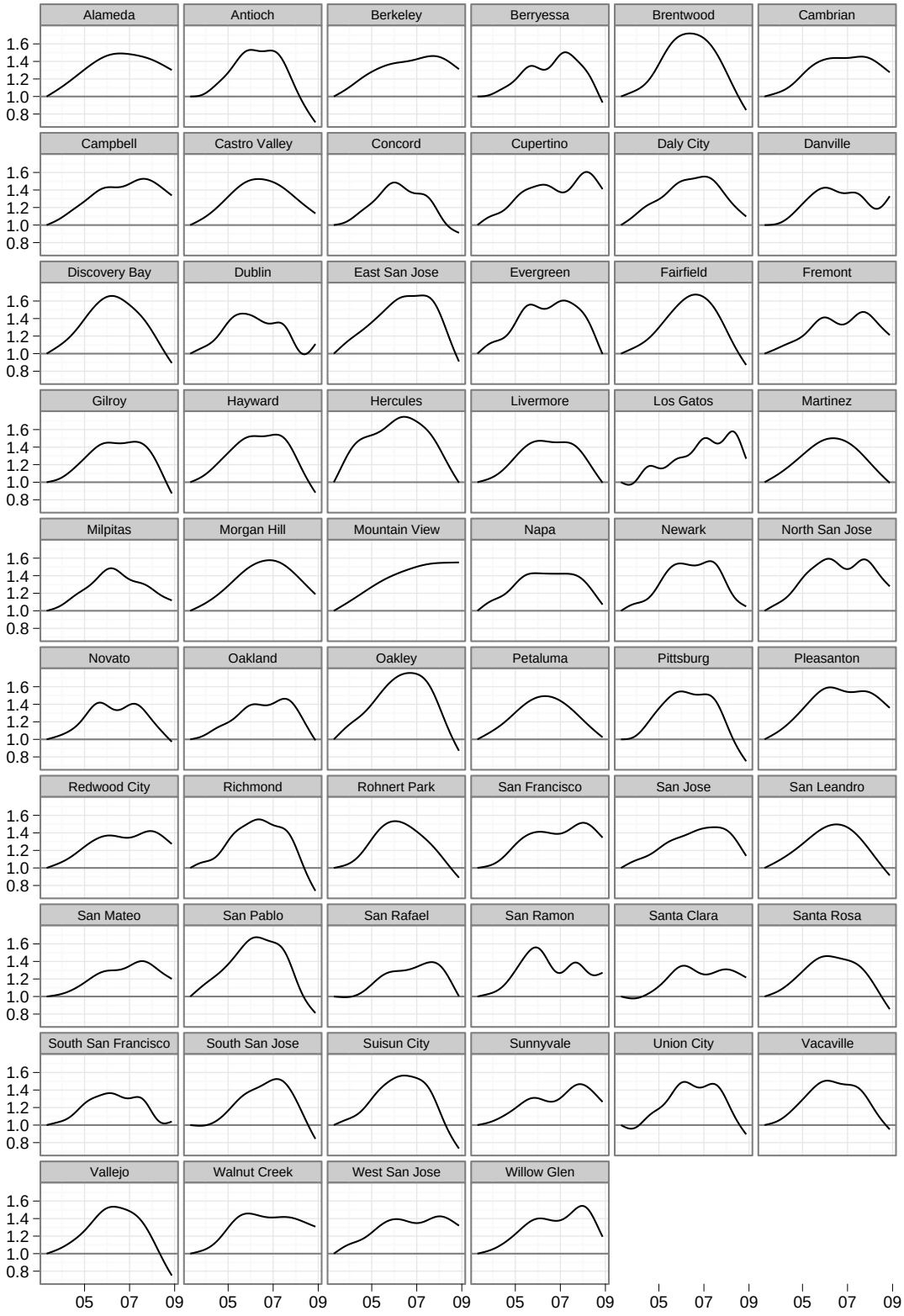

Figure 11: Individual plots for each city.

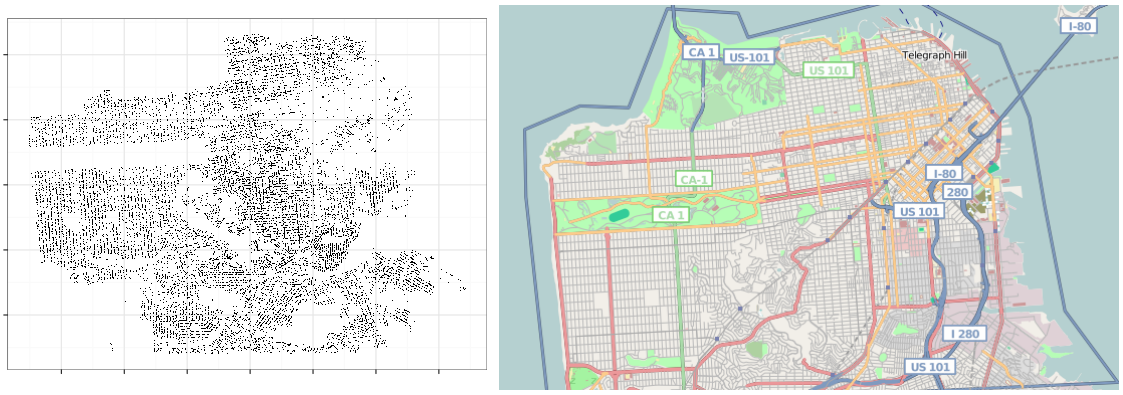

Figure 12: (Left) A small point is drawn for every house sale in the data. It gives us a pretty feel for the layout of San Franciscio. (Right) For comparison, a street map of San Francisco from `http://openstreetmap.com`

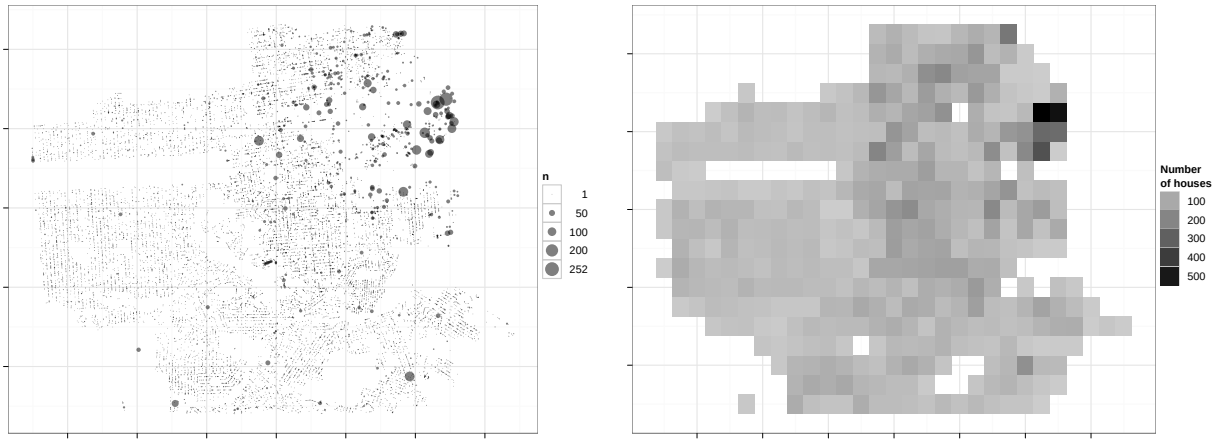

Figure 13: The geographic distribution of numbers of house sales. (Left) this plot is similar to the previous plot, but the size of the dot is now proportional to the number of sales at each unique location. This changes the picture significantly, as the large apartment complexes in the city now pop out. (Right) A display of housing sales at a higher level of aggregation: latitude and longitude are divided into a small number of bins and the number of sales in each bin is counted and displayed as the colour of the bin.

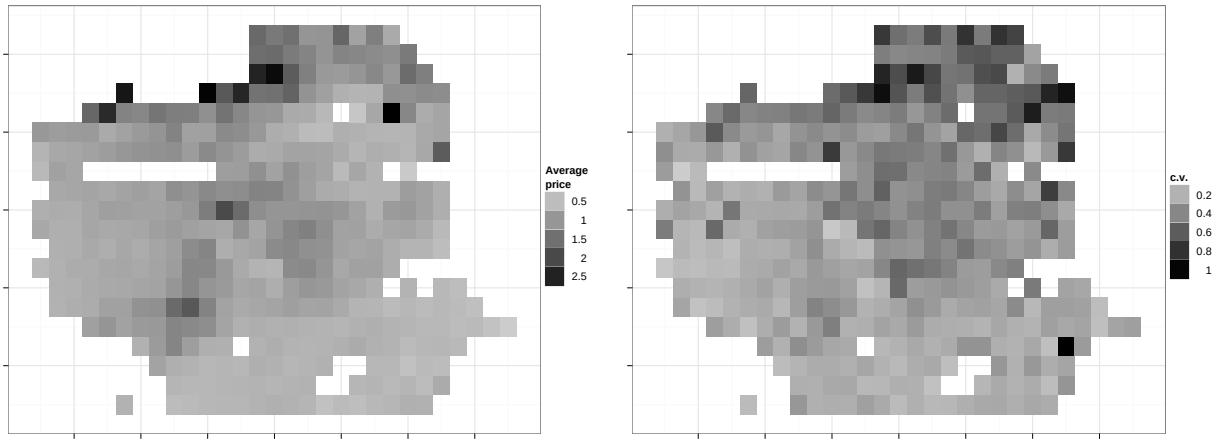

Figure 14: Geographic distribution of house prices. Using the same binning as above, the mean (left) and coefficient of variation (right) are computed and displayed using shades of grey.

to follow our work in more detail and try out some your own ideas, we have made all data and code available in a git repository at `https://github.com/hadley/sfhousing`. All the tools we used are open source: you can download them and replicate our work yourselves. This principle of reproducibility (Gentleman and Temple Lang, 2007) is very important for science: we provide enough detail that you can follow our work every step of the way, and you can run a script to reproduce exactly what we did. If we made a mistake, you can easily discover it, fix it and observe the effects on our conclusions.

Making a data analysis reproducible can be a lot of extra work, but once the principles are ingrained in your workflow, it doesn't take that much time. Importantly, it is not only useful for others, but also to future-you. If you haven't made your analysis reproducible and you come back to it 6 months or a year later (quite common if you are submitting the results to an academic journal), you may look at the code and wonder what on earth you were thinking. If you have made the analysis reproducible and included plenty of comments about your thinking, then you save a lot of time trying to reproduce your previous work and state of mind. It also very useful if your data changes. In our case, we updated just before writing up the paper so that we had the latest data off the website. Data changes a lot more than you might think. Even when your data is about something that has already happened, often the data will change as errors are discovered and fixed. Every statistician has a story about a nightmare client whose data would not stay the same from week-to-week.

Another tool that we find useful is source code control. In an analogous way to software development, this makes it easy to throw away parts of the analysis that don't work or have been superseded. If you read the code on the website you'll see that (by and large) the analysis follows a fairly logical flow. This is not how it starts off! Data analysis is a fairly creative process, with many blind alleys that don't lead anywhere, mistakes and alternative approaches. Leaving all of these in makes it hard to follow exactly what we did, but removing them completely makes it hard to see all the things that we tried. Using a code versioning system leaves us with a happy medium where the false paths are not immediately visible, but can be looked up (although current tools are somewhat lacking for this purpose).

We love working with data, exploring it and learning what is going on. We hope we have shared with you our enthusiasm for and excitement about data analysis, and have shown you some of the tools and techniques that we find most useful.

References

Robert Gentleman and Duncan Temple Lang. Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics*, 16(1):1–23, 2007.

Daniel W. Goldberg. A geocoding best practices guide. Technical report, GIS Research Laboratory, University of Southern California, 2008. URL `http://www.naaccr.org/filesystem/pdf/Geocoding_Best_Practices.pdf`.

D.W. Goldberg and J.P. Wilson. USC WebGIS Services, 2008. URL `https://webgis.usc.edu`. Last accessed December, 2008.

Simon Wood. *Generalized Additive Models: An Introduction with R*. Chapman Hall/CRC, 2006.