Bay area blues: the effect of the housing crisis

Hadley Wickham, Deborah F. Swayne and David Poole

January 12, 2009

Introduction

The housing market has received a great deal of attention in the media for the past several years. From about 2000 until 2006, we watched with excitement and apprehension as prices soared; since then, we've watched them tumble as credit became scarce and foreclosures mounted. In this chapter, we take a closer look at this story by analyzing the sales of half a million homes in the San Francisco Bay Area from 2003 to 2008. What can we learn about the way prices rose and fell throughout a single region and across a wide range of prices?

We begin by describing the data, how we obtained it, and how we prepared it for analysis by restructuring, transforming, cleaning, and augmenting the raw data. As our analysis proceeds, we communicate most of our observations using graphical displays. Along the way, we will also describe some of the tools we use, most of which are freely available. Our main tool is R, a statistical programming and data analysis environment, and we use it at all stages: fetching, cleaning, analysis, diagnostics, and presentation.

How did we get the data?

Once we decided that we were interested in real estate sales, the search for data began. Data searches are not always successful, but so we felt particularly lucky when we found weekly sales of residential real estate (houses, apartments, condominiums, etc.) for the Bay Area produced by the San Francisco Chronicle, `http://www.sfgate.com/homesales/`. We felt even luckier when we figured out that we didn't have to extract the data by parsing web pages, but that the data are already available in a machine-readable format. Each human-readable (html web page) weekly summary is built from a 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 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 this list of dates in hand, we then generated a list of urls in the correct format and downloaded them

with the Unix command line tool `wget`. We used wget because it can easily resume where it left off if interrupted.

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 is no standards document that exactly describes the structure of csv files, for most statistical datasets it is not complicated and every statistical package (and Excel!) can read them without problems. This gives us a csv file of the form:

```
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 than the previous form, but is very easy to work with using R. Another minor advantage is that it is 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. We must take care to account for the missing values in our analysis.

It takes only a few minutes to parse the files for all 293 weeks and create `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 quite common in data analysis: it can take far longer to prepare the data in a consistent, machine-readable form than it takes to perform the actual analysis.

Geocoding

When we first looked at these data, we thought it would be really important to geocode all 436,106 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. This is 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 unsuitable for two reasons: they impose strict daily limits on the number of requests, and there are cumbersome restrictions on the use of the resulting data. The request limit alone meant that it would take well over a month to geocode all the addresses, and then the licensing would have affected publication of the results! After further investigation we found a very useful open service, the USC WebGIS, 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. There was no daily usage cap when we began using the service, 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 is not 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).

As well as latitude and longitude, the USC results also include a categorical varable indicating their degree of accuracy: exact address, zip code, county, etc.

Data checking

It is in general worth spending a significant amount of time at every stage of an analysis to make sure that data are accurate, and this stage was no different. Errors in geocoding came from a number of sources: There are typographical errors in the addresses, new buildings are often not listed in public databases, and zip codes may be reassigned over time. We further suspect that the USC software included a bug during the period we used it, because large numbers of addresses were falsely assigned to the Los Angeles area and elsewhere around the state; we remapped these addresses using another free on-line service at `gpsvisualizer.com`. Our debugging process included using R to draw simple maps of latitude vs. longitude for each county and most towns to identify the addresses that had been located far afield.

The addresses in San Jose posed an interesting geocoding challenge. Sales are listed for several "towns" that are not recognized by 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.

Where possible we tried to correct any errors. When that was not possible, we used R's missing values to indicate that we do not know the exact latitude and longitude. This is a better approach than throwing out bad matches, because we need varying levels of accuracy for different purposes: When we map the data at the level of county or city, we can be satisfied with an approximate location. The use of missing values for latitude and longitude ensures that any location with a suspicious geocoding will be dropped from analyses that use latitude and longitude, but included in all others.

Analysis

It is best to start an analysis with a very broad overview. Since we are interested in the housing crisis, we will look at the weekly number of sales and the average price. Once we have a feel for the overall patterns, we will divide the data into smaller pieces and see how they compare to each other and to the overall patterns. We will examine the data using house price (from most expensive to least), and spatially, both between cities and within a single city (San Francisco).

Figure 1 shows weekly average sale price and number of sales for the 293 weeks in the data. There are some very interesting patterns. The behavior of the average price is striking, with an increasing trend until June 2007 and then a precipitous drop to the present day – a clear illustration of the boom and bust in housing prices. Sales show a surprisingly different pattern. Starting near the middle of 2006, sales volume decreases until early 2008, when it begins suddenly to increase. One possibility is that by this point house prices had dropped enough that buyers were shopping for bargains again.

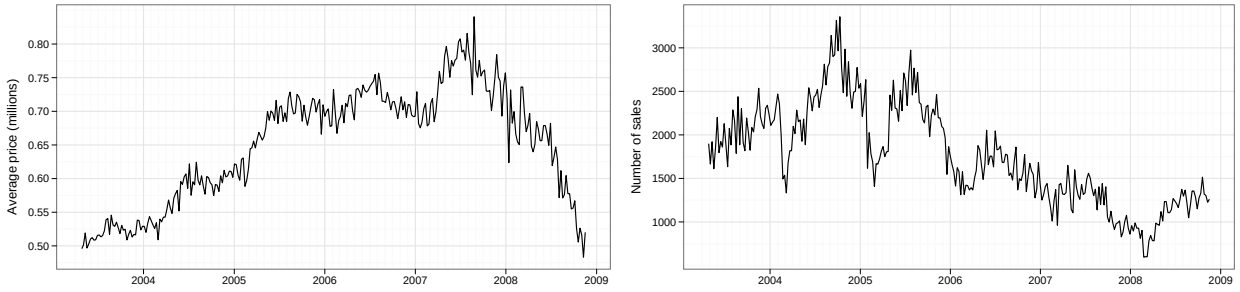

Figure 1: Weekly average prices (left) and sales (right).

The influence of inflation

The data were collected over a relatively short period of time (almost 6 years), but we might wonder if it is 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 and services. This price is calculated monthly, and we will use the West coast series, series CUUR0400SA0, to adjust for inflation as follows. We want to adjust all values to 2003 dollars, so we divide each CPI value by its value in March 2003. This operation is also known as indexing. It 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. 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 we can see that the inflation-adjusted rise in house prices is slightly less pronounced than the unadjusted trend.

Finally, though, we decided not to adjust the sale prices for inflation. Housing prices have an influence on the CPI, because one of its sub-indices is a housing index, a measure of rent and "owner's equivalent rent." It could probably be argued that housing prices had a significant effect on the CPI throughout the period under study.

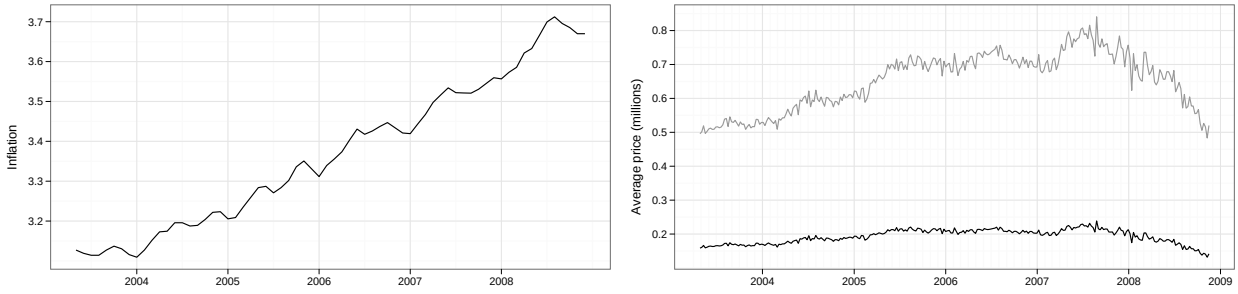

Figure 2: (Left) Inflation, indexed at 1 at start of series. (Right) Inflation-adjusted house prices in 2003 dollars (black), and unadjusted prices (grey). Failing to adjust for inflation makes the rise look a bit steeper, but has little effect on the decline. Monterey, San Benito, San Joaquin, and and Santa Cruz counties excluded because we only have data for 2008.

With this basic overview in hand, we now drill down into the details. In the following sections we break the house sales into smaller groups, first by price and then by location. We are interested in finding out whether the housing crisis has affected some groups of homeowners 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 are looking at the Bay Area, so homes will be more expensive than many other places in the country, but we still expect to see some relative inequalities. (NB. In the following, we will frequently use the word "houses" to refer to all categories of residential real estate: houses, townhouses, apartments, etc.)

As a first step, we calculate price deciles for each month. The deciles are the nine prices for which 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90% of houses cost less. This is a succinct summary of the *distribution* of the prices for each month: instead of just looking at the average price, as we did earlier, we have nine numbers that summarise the complete distribution of the prices. (We don't display the curves for

the minimum or maximum price, because they would be too choppy.)

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.

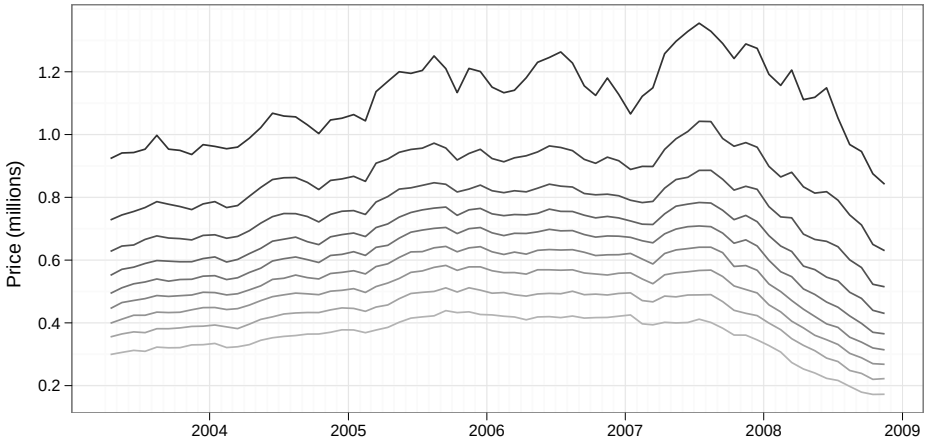

Figure 3: Monthly average house price within each decile. Lower deciles have lighter colours.

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 1.0 and we can see the relative change in price over time. The interesting aspect of this plot is that the cheaper houses (the lighter coloured lines) seem to peak higher and earlier (mid 2005), and then drop more rapidly thereafter. The cheapest houses, in the lowest decile, lost 43% of their 2003 value compared to only 9% for the most expensive houses. Comparing Figures 4 and 4, we see that although the biggest decline in actual prices occurred at the expensive end, it was the cheapest houses that proportionately lost the most value.

Another way to look at this inequality is Figure 5. Here we have divided all the prices by the median price. The values now represent a proportion of the median house price: A value of 1.2 represents a price 20% higher than the median, and 0.8 is 20% lower. Since the beginning of 2007 (before the slump began) the relative inequality has been growing. This is interesting: Does it suggest that a widening of the price gap between expensive and cheap homes is a precursor to a subsequent crisis? Has this preceded other crises? These questions could be investigated further.

Geographic differences

In this section we explore the changes in home prices in different cities in the Bay area. Because we are looking at average prices, we must take care not to include cities with only a few sales. We decided to focus on all cities with an average of at least 10 sales per week. This gave us 58 cities (24% of the 245 cities in the data) with 428,415 sales (82% of the sales).

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 display: the spaghetti plot. It's

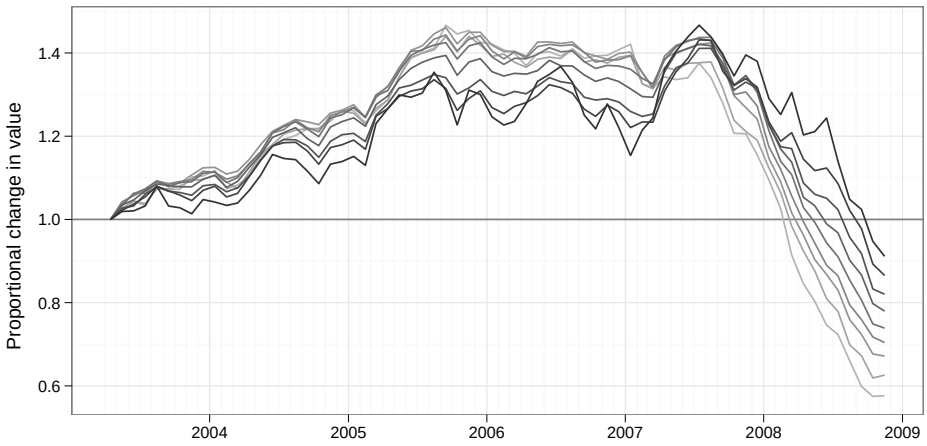

Figure 4: Indexed house price within each decile. The average price of cheaper houses peaked higher and earlier, and fell more steeply.

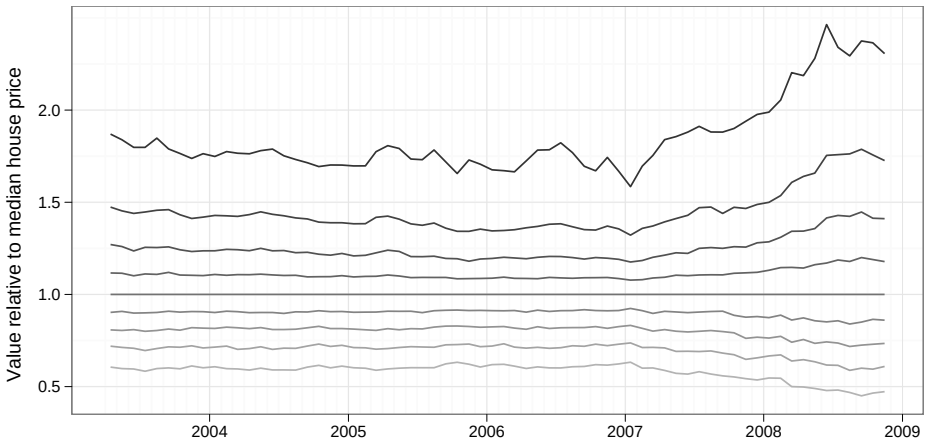

Figure 5: House prices, relative to the price of the median priced home. The disparity in home prices is increasing.

very hard to see anything in the big jumble of lines. One method of improvement is to smooth each line, removing short-term variation and allowing us to focus on the long-term trends we are looking for.

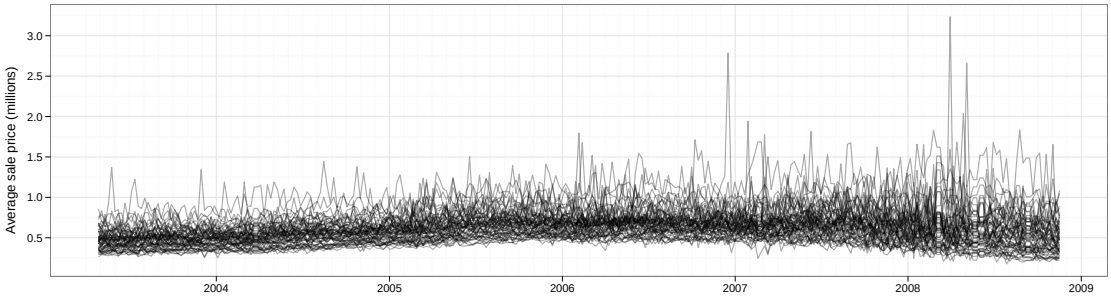

Figure 6: Average sale price for each week for each city. This type of plot is often called a spaghetti plot.

To create smooth curves, we used generalised additive models (GAM), a generalisation of linear models (Wood, 2006). This method fits smooth curves by optimising the trade off between being close to the data and being very smooth, in effect removing noisy short-term effects and to emphasizing the long-term trend. This is exactly what we need: we are not interested in daily or weekly changes, only the long-term changes related to the housing crisis.

The left part of Figure 7 shows 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 a few very expensive houses. We will also index each city in the same way we indexed each decile: dividing by the starting price puts each city onto a common scale and allows us to focus on the changes. This is shown on the right of Figure 7.

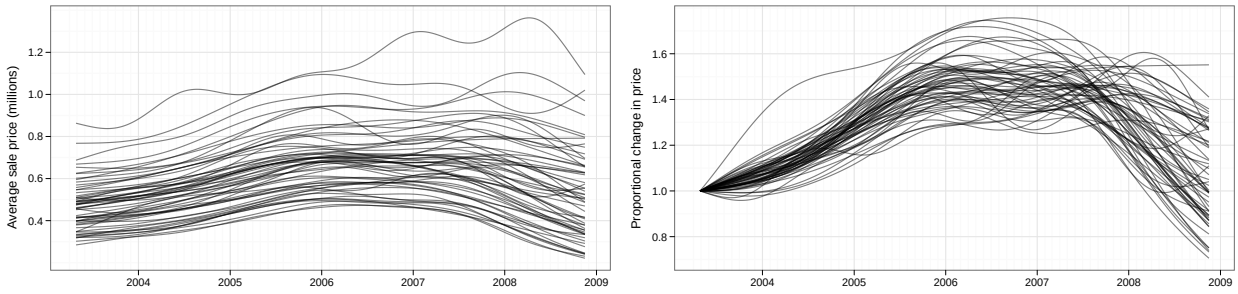

Figure 7: Smoothed city-level weekly average sale prices. Compared to the non-smoothed version it's easier to see the long-term trends, but it's still not particularly easy.

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 8. This plot takes up a lot of space but is worthwhile for the extra information it affords. We can pick out some interesting patterns: Berkeley and San Francisco show less of a peak and less of a drop, and Mountain View is unique in that it has seen no drop at all in housing prices. Other cities such as Oakley, Vallejo, and San Pablo, show both big peaks and big drops.

Recall that in our earlier discussion about San Jose, we noted that the raw data describes many neighbourhoods of San Jose as cities in their own right. We know that this data is a bit messy, with the same address occasionally being associated with different neighbourhoods, but this data suggests that the neigh-

bourhoods have distinct characters. Berryessa, East San Jose, North San Jose, and South San Jose have similar curves, showing a sharp peak and an equally sharp drop; Cambrian, San Jose, and West San Jose, on the other hand, don't show much of a decline.

After further investigation we concluded 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 created a new variable called *price drop*, which is the difference between the price in February 2006 (at the height of the boom) and the price in November 2008 (the doldrums at the time of writing). Figure 9 groups the cities by this new variable. The divisions are arbitrary, but one can see how the cities in each group follow a similar pattern: the bigger the boom, the bigger the collapse. This suggests that this single number does a good job of summarising the different effects of the housing bust.

We have determined that cities have different patterns, but we don't yet know why that might be so. The geographic pattern, as in Figure 10 does not reveal anything particularly striking except that the worst hit towns tend to be to the north and the east. This does not offer much in the way of explanatory power, so we looked for additional data that might help us gain a deeper understanding.

Census information

The US Census Bureau provides demographic data from recent surveys at both the county and city levels. The quickfacts website, e.g., `http://quickfacts.census.gov/qfd/states/06/0649670.html`, displays a number of interesting demographic variables for each city. Unfortunately, city-level data are not available in an easily downloadable format, but we were able to use scripting methods (like those we used for the sales data) to collect the demographic information and convert it into csv. In addition, the definition of a city differed slightly between the census data and the sales data, so we could only match 46 out of the full 58 cities. The census data didn't cover some of cities we chose because their population was below some cutoff, and some of what the housing data calls "cities" are actually neighbourhoods within larger cities, as we noted earlier with respect to San Jose.

A glance 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 many of the other relationships. Figure 11 includes three scatterplots that illustrate the relationship between the drop in home prices and income, percentage of college graduates, and commute time. The correlation between *price drop* and commute time is weak, but note that all of the cities with the longest commute times (more than 35 minutes) have particularly large drops in price. It appears that the housing crisis has been relatively more damaging in poorer areas.

The county-level census data contains more variables than the data for cities, so we analysed the county data for further explanation of the housing crisis. The plot on the left in Figure 12 shows, for each county in which we had sales data, the percentage change in the number of housing units (from 2000 to 2006) plotted against the median sale price in 2008. There is a strong negative relationship between recent home values and the amount of new construction. In other words, most of the building boom in recent years occurred in poorer neighbourhoods, and as we noted above, these are also the areas where the subsequent slump has been the most severe. San Joaquin county, in particular, which has consistently low prices across towns, experienced by far the most new construction in recent years. We should note that we do not have many sales in a few of these counties (e.g., San Benito and Santa Cruz), but the overall nature of this relationship is still very clear. The effect is further illustrated in the right-hand plot in Figure 12, which shows the median price for all the sales in each year, as well as the median sale price for the subset of homes built in that year (and then sold in that year or later). We see that despite the large overall boom from 2003 to 2006, the new construction during this period was aimed increasingly at the lower end of the market.

According to an article in the New York Times (McKinley, 2007), the city of Stockton, one of the larger

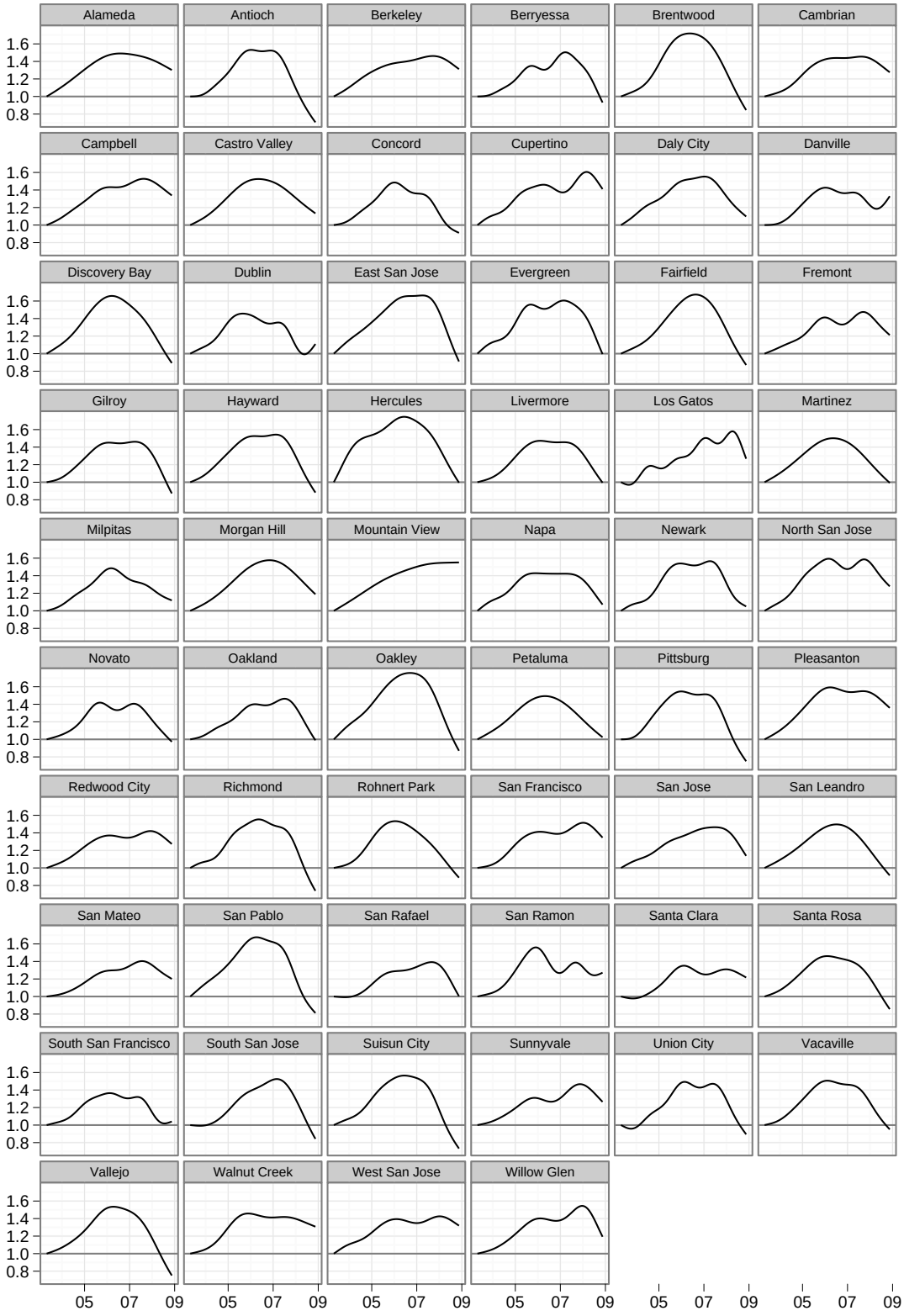

Figure 8: Individual plots for each city.

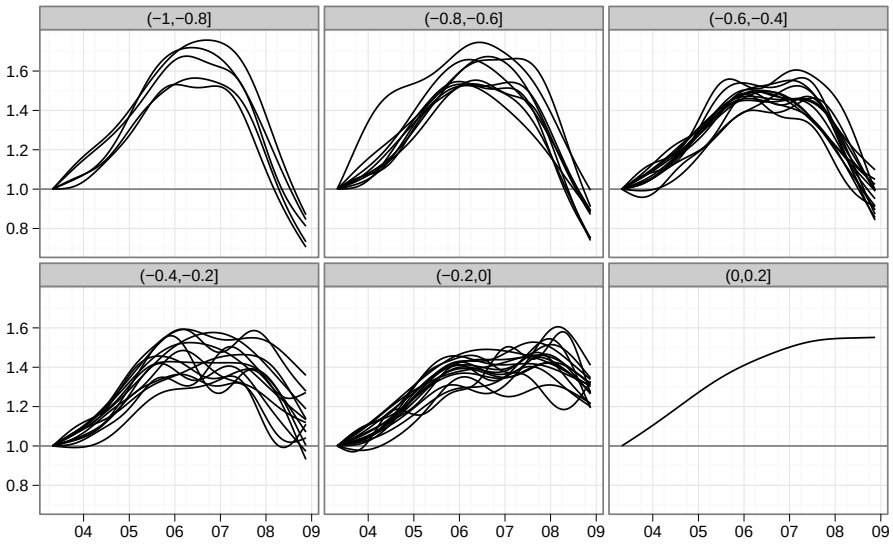

Figure 9: Plots grouping the curves for towns by their value of *pricedrop*. The towns in the upper left plot had the largest price declines (between -.8 and -1); the town at the lower right (Mountainview) is the only one that shows no decline. The patterns within each group are similar, suggesting that this single number provides a useful way to divide the cities into groups.

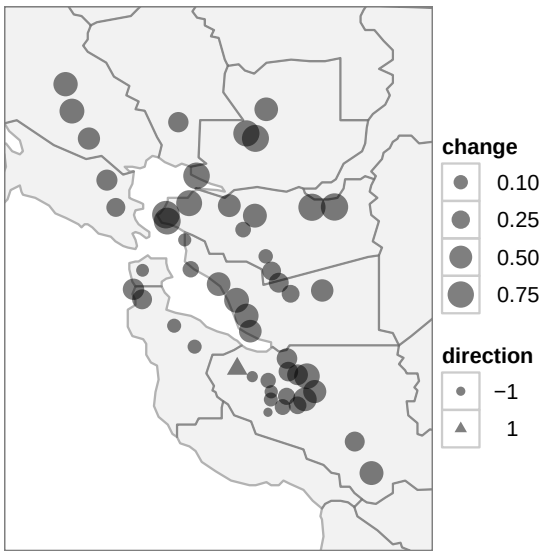

Figure 10: The geographic distribution of price drops.

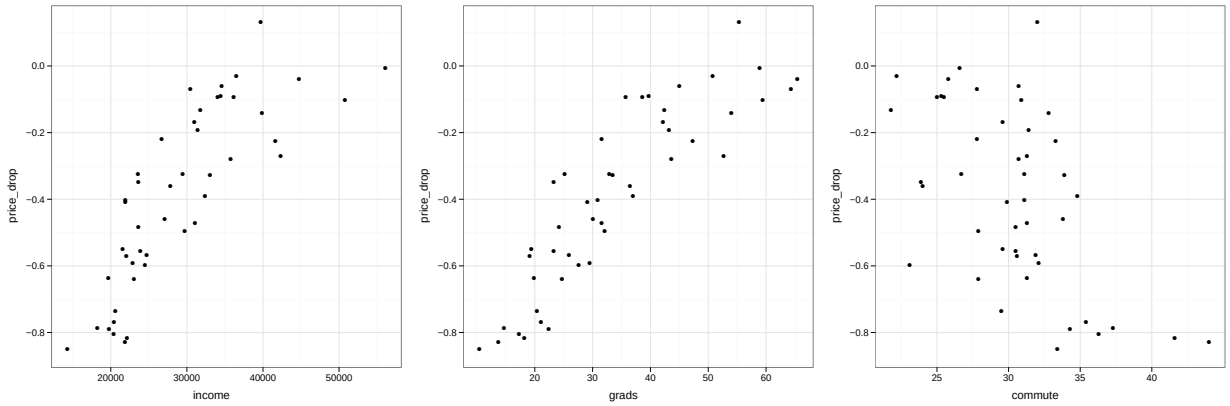

Figure 11: From left to right, the relationship between the drop in house prices and average income, percent of college graduates and average commute time.

Figure 12: Influence of new construction on recent prices. [to be regenerated]

cities in San Joaquin county, already had the highest rate of foreclosures in the USA by the summer of 2007. Unfortunately we do not have any sales for Stockton prior to 2008, but it appears it was a leading indicator of the slump in the region that would continue into 2008. The population of Stockton grew rapidly in the last decade as commuters moved further out to escape the overheated housing market in the immediate Bay area. This helps to explain the new construction noted earlier, and also ties into our observation regarding commute times. The article also lists Modesto and Merced, two other towns in the Central Valley, in the top 10 nationwide for foreclosures at that time.

Exploring San Francisco

Having explored the difference between cities, we turned to look at a single city in more detail. San Francisco is the obvious choice: It is the largest city in the data, it is the city with which we are most familiar, and it has some iconic features that should be easy for others to identify as well. We started our exploration by extracting all addresses within San Francisco that were geocoded with a fairly high degree of accuracy, giving us a total of 25,377 addresses. We created a simple scatterplot of the latitudes and longitudes, Figure 13. For the residential parts of the city, this gives an amazingly detailed picture. We can see the orientation of the streets, the waterfront boundaries and parks. Our view of some areas, like downtown, is more patchy because there are fewer residential homes there. (In this section, we will avoid using the shorthand term "house" since it is obvious that so many of the home sales represent apartments.)

One problem with this plot is we cannot see the number of sales at each specific location. Figure 14 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 many sales there. Looking more closely at the data reveals that these are apartment buildings with hundreds of apartments. On the right, we have divided San Francisco into squares of 0.005 latitude and longitude and counted the number of homes in each bin. This gives us a higher level view showing where the majority of homes are located.

Using that same binning, we calculated the mean and coefficient of variation of the home 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.

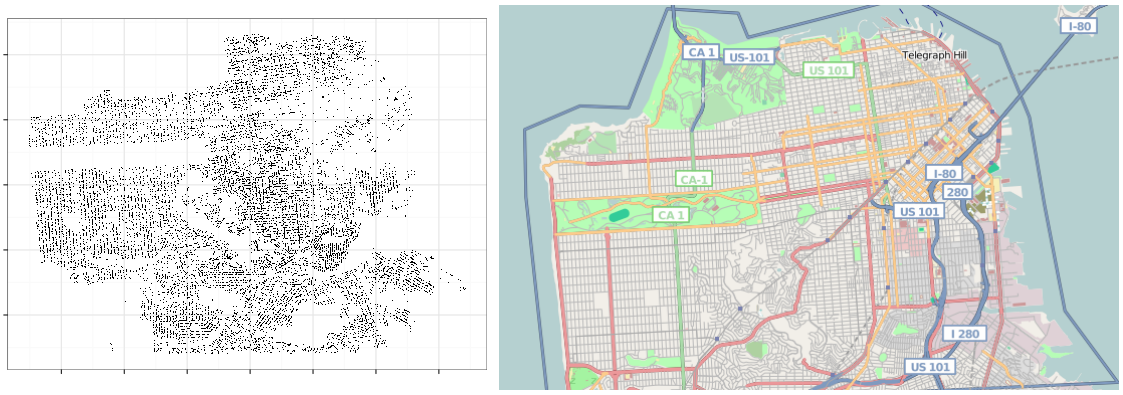

Figure 13: (Left) A small point is drawn for every home sale in the data. It gives us a pretty feel for the layout of San Franciscio. (Right) For comparison, a street map of San Francisco from `http://openstreetmap.com`

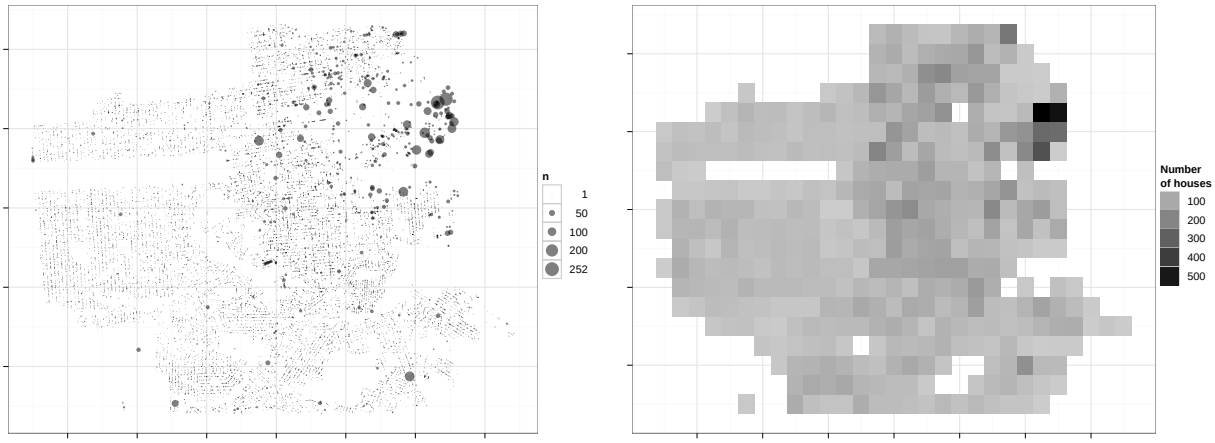

Figure 14: The geographic distribution of numbers of home (house, apartment, etc.) sales. (Left) this plot is similar to the previous plot, but the size of the dot is now proportional to the number of sales at each unique location. This changes the picture significantly, as the large apartment complexes in the city now pop out. (Right) A display of home 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 15 shows the geographic distribution of these two summary statistics. We can see the most expensive homes border the Presidio and coast to the North of the city. 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 appears to increase towards the Northwest.

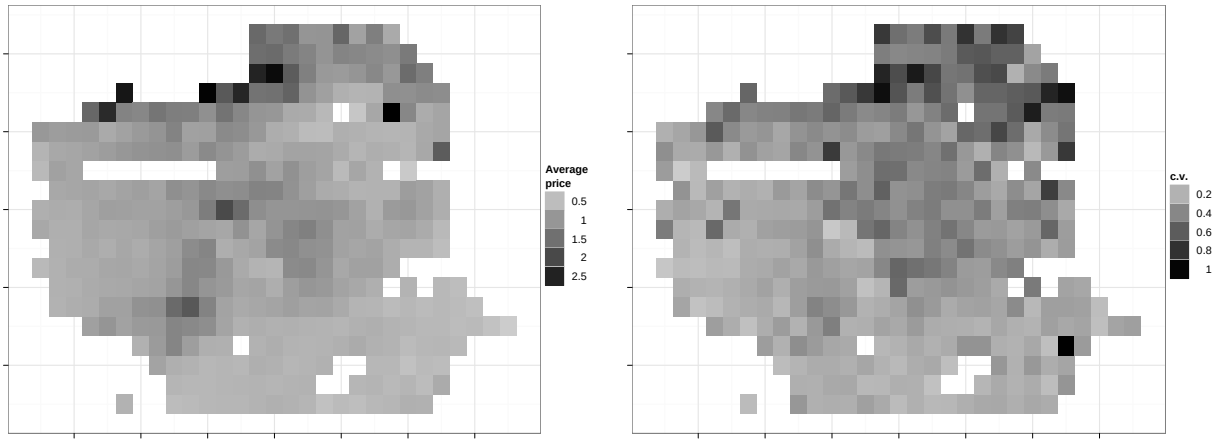

Figure 15: Geographic distribution of home prices. Using the same binning as above, the mean (left) and coefficient of variation (right) are computed and displayed using shades of grey.

Conclusions

WE SHOULD START THE CONCLUSION WITH A SUMMARY OF THE FINDINGS ABOUT THE HOUSING CRISIS. CAN BE JUST A SENTENCE OR TWO.

We have used relatively simple statistical tools such as indexing, quantiles, smoothing and binning to explore a large and complex data set. We began with broad summaries and then dug deeper to explore the details, but we have only just scratched the surface. If the data has caught your interest and you'd like to follow our work in more detail and try out some your own ideas, we have made the all data and code available in a repository at `https://github.com/hadley/sfhousing`. All the tools we used are open source: they are available for download and our work can be replicated. This principle of reproducibility (Gentleman and Temple Lang, 2007) is very important for science: we provide sufficient detail that you can follow our work every step of the way, and you can run code 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.

The creation of a reproducible data analysis can be a lot of extra work, but once the principles are ingrained in the workflow, it doesn't take that much time. Importantly, it is not only useful for others, but also to the original analyst, who in time will likely forget the finer details of the work. A well-commented reproducible analysis will save a lot of time when revisited later.

Another tool that we find useful is source code control. In an analogous way to software development, this makes it easy to discard parts of the analysis that led nowhere or have been superseded. The code on the repository website indicates 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, mistakes and alternative approaches. Inclusion of all of these makes it hard to follow exactly what we did, but removing them completely makes it hard to see all the things that we tried.

We enjoy working with data, exploring and learning from it. We hope we have shared with you our enthusiasm for 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.

Jesse McKinley. From housing haven to foreclosure leader. New York Times, August 13 2007. www.nytimes.com.

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