CURS Symposium Narrative

Parker Kemp

Slide 1:

Title Slide

Slide 2 and 3:

Text is in slides.

Slide 4 and 5:

Our first task was selecting a set of counties across the US from which to gather data. The motivation behind their selection was to choose counties that were from as different geographic regions as possible, had similar population sizes, and had readily available data that we could access and use. The reason that we needed to select counties from different geographic regions was so that we get a diverse representation of pool ownership in the US, because we believe that pool ownership will differ by region due to climate.

We then created a random sample of 4,000 addresses that were coded as ‘Single Family Home’ for each county, which would be the samples we collect data on for the remaining portions of the project. It is important to use only ‘Single Family Home’ addresses, so that we can directly associate a pool with a given address, rather than something such as an apartment complex where many apartments may share a community pool.

Slide 6 and 7:

The 3 data collection approaches we considered are outlined here.

First, our original data source for the project, web scraped data. From the sample of single family homes that we created, we realized that in order to web scrape the data that we wanted, we would need to obtain URLs for each of the websites we were interested in scraping. To achieve this, we wrote a script that performed a google search of each address plus the keyword ‘Zillow’ or ‘Trulia’, and gathered the first URL returned on the google search page. We then had the task of extracting the address from the returned URL and checking to see if it matched our original searched address. Those addresses that matched would be saved to later be run through a web scraper that would gather property details from each address’s listing on the real estate sites.

Another data source considered is vendor curated data. We provided a vendor, Marketing Systems Group, with a spreadsheet of our sampled addresses, and they returned data on each address that they were able to match.

Finally, we are considering data gathered through Zillow’s API, which is an easy and intuitive way to gather data for a given address, if Zillow contains data on that address.

Slide 5 illustrates the high-level data collection strategy for clearer understanding of the data flow and process.

Slide 8:

With the URLs from the google scraper returned, we needed to see how many of those URLs were for addresses matching the address we had searched, and how many were errors. To do this, we parsed out the address contained in the URL, and compared it to the address that we had searched, using exact matching. This means that addresses that may actually be the same but have a spelling error or formatting difference would not be matched. Due to this, we believe there is some bias in our match rates, particularly in Thurston county, as they format their addresses differently from the other counties – which leads to fewer matches than we would otherwise have. However, to maintain an efficient, understandable, and repeatable process we chose to use exact matching rather than a more complex technique.

If we wanted to improve the overall match rate we could implement a more complex method such as standardizing every address to the same format before matching or performing fuzzy matching, in which we would call two addresses a match if they contain a certain percentage of identical characters, even out of order. We did not explore these methods here, but they are an area of interest for future exploration.

Slide 9:

Here we can see that across all four counties, Zillow matched a higher proportion of the sample than did Trulia, and we got a consistent number of matches across counties. We believe this may be because Zillow is likely a more popular real estate platform, and thus has more robust data and contains more address listings than Trulia.

We expected that Thurston have a lower match rate due to peculiarities in their county data address formatting, which is reflected here.

We can see that if our primary interest is collecting data on as much of the sample of interest as possible, it is ideal to use Zillow as a data source as compared to Trulia. However, other considerations must be made, such as the completeness and variety of data available from each source – which will be explored in the next phase of this project and is not reported on here.

Slide 10:

Here we compare the match rate for all four data sources. You can see that the Zillow API outperforms all other data sources for every county.

There appears to be some consistency for each data source across the counties, with the exception of the vendor. The match rates for the vendor vary quite a bit, which we attribute to the composition of the counties: Lorain county, for example, is more rural and may have fewer addresses in consumer databases such as Marketing System Group’s.

Slide 11:

Here we look at the availability of features from each data source, another important consideration in addition to completeness of the sample. Each larger region contains all of those variables contained in the region inside it, in addition to the variables exclusive to that data source.

We can see that the Zillow API does not give very rich data, returning only a small set of features, especially compared to the other data sources.

When web scraping from Zillow.com or Trulia.com, we will be able to obtain a much larger set of features. This includes those listed for the API as well as the additional features listed here. Because we can access any of the data available on the listing site, there is much greater flexibility in collecting exactly the data we are interested in, and we are not restricted by what Zillow is willing to provide us with.

In terms of data variety, the vendor data is much more complete, returning over 1,500 variables. However, we must consider that these data were purchased from the vendor, while the others are obtained for free through other means. We must weigh the benefits that come with the greater variety of data the vendor provides with the costs of obtaining it.

Slide 12 and 13:

These slides analyze the consistency between the Zillow API data and the vendor data for Alachua county. There 5 variables are those that both data sources have in common, allowing us to perform a fair comparison.

Each plot shows the distribution of the absolute difference in value between the API data and the vendor data, in other words, how far apart the values are for a given observation in each data source. Ideally, we would like to see plots concentrated at 0, indicating that the data values correspond to each other exactly, and there is no difference between the data sources. We see this is some cases, such as year built and number of bedrooms, but unfortunately this is not the case for other variables.

The distribution for number of bathrooms is centered at 1 and indicates that there is some significant error in the reporting of the number of bathrooms between the two data sources.

Likewise, the distribution of home square footage is centered around 25 in Alachua, and 12.5 in Anoka. In this case, the vendor provided this data as an ordinal categorical variable (a value of 11 indicates square footage of 1,100). Because of this, we must convert the API data, which was continuous, into a similar format. Therefore, the value of 25 that the distribution is centered around, is indicative of a 250 square foot difference.

These are significant differences in the property information, and we would need to perform a deeper exploration to confirm the accuracy of each data source. But we can confirm that there are differences between the vendor data and the open sourced data from the API, so we have to be careful about how we use these data for modeling.

Slides 14, 15, 16:

Text is on slides