Session 1: Thursday, September 14th 3-4pm

Location: Ambassador I

Session Title: Rental Use Market Trends and Data Scraping

Organizer: Jessie Partridge

Primary Notetaker: Maia Woluchem

Present: Stephanie Rosoff, Ashley Williams-Clark, John Kileen, John Cruz, Liza Morehead, Jeremy Pyne, Nancy Jones, Brian Herron, Cherie Chung, Randy Rosso, Katy Getsie, Eamon Johnson, additional attendees

**Some Resources:**

MAPC Rental Listings Scraper Github: <https://github.com/MAPC/rental-listing-aggregator>

MAPC’s Dimensions of Displacement: Managing Neighborhood Change report: <http://www.mapc.org/managing-change-somerville-dimensions-displacement-report-released>

Craigslist and U.S. Rental Housing Markets, Geoff Boeing, UC Berkeley: <http://geoffboeing.com/2016/08/craigslist-rental-housing-insights/#more-1930>

Stephanie Rosoff – We want a rental index. I think we’re going to get it.

Ashley Williams-Clark – I do IDS work and work on affordability and homelessness. Charlotte’s had a huge boom in rental development. We did work on rent as well and what’s been in charge in this development.

John Kileen – Work for the City of Durham right now. Working on developing metrics around neighborhood change and gentrification. One of the biggest gaps is real time rental info. And not just where rents are rising but also where they’re lower already. I’d like to see both sides of that and get to things that are consistent.

John Cruz – With Rise Community Development. We’re primarily a homebuilder. Right now we ask our CDCs what the market rate is in the area. I’ve looked at the Boston tool and I think it’s something that we can really use. I’d like to have real numbers to back up what we’re saying.

Liza Morehead – From Portland. We’re getting ready to dig into an initiative to create a comprehensive housing database and portal. Rental data will be an amazing thing to be able to add to that.

Jeremy Pyne – Grand Rapids is in low supply of rental units. Hard to find rental, let alone affordable rental. We’re seeing $1500/$2000 rentals going up next to houses. More landlords are starting to figure that out and now they’re hiking the prices. A lot of organizations over there are trying to fix it and control it.

Nancy Jones – With BNIA. We have a lot of housing data but we don’t have rental price. We know what units are rental units but we don’t update it quarterly or anything like that. But we just have a vacant housing problem. The housing voucher program is not abused but in the depressed neighborhoods, housing vouchers are active and the landlords are charging the regional market rate. So that inflates the rental cost of really bad units in economically depressed areas. We’re not advocating professionally but personally we are. We want to get rental rate data.

Brian Herron – Same story. Not a lot of supply, lots of demand. I’d like to get an idea of where it’s rising more rapidly.

Cherie Chung – Work at Digital C in Cleveland. Trying to build an open data portal for Cleveland. We’re also doing work on land-use and blight. We’d love to learn more about how to include rental rate data into the indicators and the apps for those counties.

Randy Rosso – Citizen of Arlington, VA where rental is a major issue. Arlington has made the real estate assessments data open but.

Katy Getsie – With the MetroLab Network. A previous citizen of Pittsburgh. Now we have a lot of new development. The markets are all over the place. Would be really interesting data.

Eamon Johnson – I don’t know a lot about rental units but I know a lot about scraping.

Jessie – We need to decide if we’re ok with tweeting. Anyone opposed to tweeting? I’m also supposed to set a goal. I want to share what we’re doing because I think it’s exciting. Who’s collecting now?

John C. – We collect from the properties that we own.

John K. – I get limited access to the MLS rental data. I can’t publish it but we also get ACS data.

Jessie – Anyone not familiar with data scraping? No? Ok. I can tell you what we’ve done and are currently doing. Our starting point with trying to get rental listing data is trying to write a report on the extension of our transit lines into Somerville. There’s one transit stop in the city but this line is running through the middle of that city. The city and our community partners are interested in understanding how it might affect displacement where the transit line was built. So we’re at the point of estimating displacement. A big part of that is rising rents. We didn’t have any sort – we had transaction data and assessors’ data. Helps with property owners but not renters. We found a local data enthusiast who had been scrapping Padmapper and publishing heat maps of rent throughout the region. Published his code on GitHub, and we were able to get listings for two years of data. And we used that to create a model and project estimated rent increase with the building or extension of this transit line. A lot of interest in some of that data. We didn’t have time or capacity to continue with that so flash forward to November of last year, we updated our own scraping tool. We scraped Padmapper and Craigslist one time a week and collect all the listings. Basically now we have a million records from Craigslist and Padmapper.

One is a legal question. There have been a number of threads on the NNIP page for this. And a lot of threads for researchers for this. So the legal issues are a big one. We have a project starting with Boston and 4 other cities. So we’re going to be moving forward and we need to justify that legally. And we have to clean that. We can remove duplicates between Craigslist and Padmapper. But within that data, there are duplicates of all different kinds. If someone posts their listing 3 times with three different exclamation points, that’s not as easy to identify and remove. There’s a bit of machine learning that we need to catch up on. There’s a spatial challenge. On Padmapper, it’s good special data. But on Craigslist it’s not clean. If they put their address it’s ok. But without an address, it places the location in the middle of the zip of city. We need a way either a way to remove those or figure out where the location is. And then a combination of spatial issues or duplicates. Basically, the deeper you look into the data, the more you see outliers. If I look at the block, I see a handful of questionable listings. Another challenge is the difference between someone posting the room and then an actual full unit. The reason I’m excited about working with the cities is that they have specific use cases. They can find the duplication patterns more quickly and give us a way to clean the data. One example is the city of Boston. They have a number of programs to control the rapidly rising rent in Boston. But they have no way to quantify whether their strategies are actually working. With that, I would open it up to folks.

Ashley – What percentage of the rental properties do you think you’re missing with this process?

Jessie - We have a sense that the listings that we are able to collect are skewing to higher market rates. I think this is – They did a project where they could show there’s less representation online in lower income neighborhoods. I don’t know if we can address it with this method. It would require surveying it in these neighborhoods.

John C. - It prompts the idea of other sites. Two other sites subsidize other rental units. How adaptable is your process to other sites?

Ashley – SocialServe – it’s almost like a craigslist for affordable housing. Works with housing authorities. You can enter in if you have a voucher, or other incomes.

John C – I’ve dug through the code you’re using a little bit and you can change the website you’d like to select. But if you want two cities, you would have to kind of start from scratch, especially if they don’t have an API.

Jessie – It can be done, but it does require more of an effort but more knowledge.

John C – Yeah, you can find it. It’s written in Ruby. And what do you know about how to get it up and running? Could I not run it on top of a VM?

Jessie – We post it to a PostGres database. It may run on its own machine. My colleague Matt runs it. It’s fairly intensive but it happens one time a week.

John C – The length is going to vary city to city.

Jessie – We do our full region so it’s big.

John C – We can’t do just the city because we are separate from our county. Lots of our county places that are unincorporated – it’s hard to market housing in Ferguson so you just say St. Louis. It works and makes sense from the standpoint of what it is, but if you say you live in St. Louis you could be in the city or in the boonies.

Nancy – I would like to know the exact process. So maybe that would warrant something more than a phone conversation?

Jessie - A big part of doing this is our documentation. We’re totally open to having phone conversations. But in terms of efficiency, it’s fun to dig through the code. It’s a challenge that we all share. They’re time and resource intensive so we haven’t had time to do documentation.

John C – Even if you could do a configuration file that would be really helpful.

Eamon – Is Craigslist – is it amenable to giving data for research?

Jessie – What we had in terms of the legality problem – **do you think it would make sense for NNIP to approach Craigslist and say that as a network we’re interested in this data for research purposes?** Because that way you come with more clout and it’s not kind of own

John K – Was there ever a point that after georeferencing and aggregating all that stuff that it had accumulated intellectual property, at which point it’s not the original data?

Jessie – We didn’t think about it at all for the Somerville project. We would like to be able to share – are we even allowed to get the data?

John C – And then would you own the statistical process and the like?

Jessie – And we want to be able to share the data in whatever way’s most useful to our municipal partners. What makes sense for summaries and what other data might be useful to join to it data such as ACS data. But there might be some municipalities that want the point-level data, so who knows.

Randy – I wonder if and how it meshes with NLIHC. They have some data that are buildings with Section 8 vouchers and the like

Jessie – We have the subsidized housing inventory that we use a lot that’s similar. I don’t think it does include vouchers but even at the local level it’s not completely accurate. Even when we came to Somerville, we used that data but they came to us with updates.

<Unknown> - Preservation of Affordable Housing Research Center. HAI is the overarching entity and this is their research group. HAI provides insurance to housing authorities. This is their research arm. I didn’t realize these were all connected until very recently.

Jessie – One other thing that we’re interested in is historical data. We have almost a year now of data. Our housing crisis started way before that. We’re interested in seeing that to see if we can get hold of data. The Way Back Machine, hosted by Archive.org might be a way.

Question for Jessie – Do you know when something is archived?

John C – Like when?

Jessie – If we pulled Craigslist from the Way Back Machine to get around the Craigslist legality issue?

John C – Websites have this thing when you can tell when a website is going to be updated. I don’t know if they take something like that into account but I don’t know if you’d be able to see data from ten years ago or something like that.

Nancy Jones – What other things to match it with…I would link it to the assessor’s data that we have but also the foreclosure data. It wouldn’t surprise me in the least to…

Jessie – Where do you get foreclosure data?

Nancy – From the court. We had a student write a web scraper for us. But the way that the scraper pulls it is that each foreclosure filing is a unique identifying number. We tried to do it for the region but everyone codes it differently. When we get it from the housing department, it’s not as thorough as when we get it from the courts. Doesn’t include the status of the filing, it just has to be filed. We used to get it through ratified sales but the time was a lot and we didn’t have the funding. We just do it when we have a student.

Jessie – We would have to go through the courts to get eviction data, but even that is paper records.

John K – The fair housing office. But low income is not a protected class. We get calls from people who have been kicked out of their place because the rent has doubled. There’s no recourse for them. Where that comes with the change in the rental value itself…

John C – All that eviction data is going to be completely incomplete no matter what. They’re going to find…

Eamon – Could you get operational sheriff’s data?

John C – If it hits that point yes. But if the landlord just says you got to go, you’re going to be out of luck.

John K - It wouldn’t get flagged anywhere.

Jessie – So what kind of specific questions would you like to be able to measure with this data? How real-time is needed? Or most appropriate?

John C – I think quarterly is sufficient. Monthly maybe, but quarterly for sure. I can’t imagine weekly things because that takes in other factors into account. During the fall, I’m sure you see the cluster of universities.

Jessie – Definitely seasonal changes I’m thinking. Not for scraping, but for reporting. So you can collect weekly and then do a quarterly analysis of it.

John C – You could – It would be easier if I still live in Montreal because 90 percent of the leases begin and end on July 1st.

Nancy – There’s got to be use variation in rental across Baltimore. That in itself would be extremely useful to tell a story – coupling that with the ACS.

Jessie – With the ACS we’ve thought about a few different attributes to join. Just joining to the median rent listed in ACS, which we know will be lower than the true rent. But it will give us a sense of how closely the data is tied to the patterns of the data. Or tie to the places where our data is represented.

Stephanie – We used the PUMS data to look at rent in the last year.

Jessie – PUMS migration data for us got really bad when they changed the boundaries.

Nancy – There’s six PUMAs in Baltimore. Within each PUMA we know there’s too much variation. We can use it for other purposes but not neighborhood scale.

Jessie – What else do people want to know?

Nancy – How to do it!

John K - The MAPC GitHub site?

Jessie – Yes.

Nancy – You asked about data cleaning. Especially with your example of the exclamation marks. We had a nice code of a computer scientist. You write the script that pulls out the special characters, and reaggregates and then doesn’t consider them.

Eamon – Recently we’re starting to look at an open source package called ChoiceMaker. USC is using this for deduplicates, the NY Dept. of Health is using it, and it has a trainable machine learning part that will tell you what fields are important and it has a sense of rules. We’re going to be talking to people from USC to see how they’re deduplicating Craigslist. There’s also a researcher at Microsoft that contains a list of deduplication materials. I found that from a link on the DataMade GitHub. There’s a DataMade in Chicago. They have a dedup tool that links to a guy in a PhD.

John C – It would be really helpful.

Ashley – When you scrape, does it capture a single-family home and the like?

Jessie – We hope to join the records to our parcel database.

John C – It should be easy enough. A massive geolocator but it’s easy in theory.

Jessie – On the point of neighborhoods – One thing about Boston is that our neighborhoods are fluid. How important is it or how important you feel it is to do summaries at custom geographies. Do you have municipally defined neighborhoods? The city of Boston has at least 3 different sets of boundaries. And none of them match. So it’s a perennial challenge for us.

Nancy – We have strict neighborhood boundaries. People may or may not identify with that but that’s how it is. But for us back in 1999 when we were starting to form, they came up with the 55 areas. We developed CSAs that our census tracts nest in. So instead of 278 neighborhoods, we have 55. So that’s how we publish it and that’s how we can put out longitudinal stuff. Even if they can’t identify, they at least can identify where they live.

John K - You might just do aggregations by blocks. We might use block groups, but it’s still masked a little bit and it’s aggregated some –

Nancy – Just internally to do the analysis. At the smallest area, it might not even be fruitful if the variation isn’t that great.

John C – St. Louis has really strictly defined neighborhood boundaries. We need to look in those neighborhoods to see what those are.

Katy – 2 questions. When you’re talking about aggregating – are you providing averages, variation, those kinds of things? And also, if you’re looking between two neighborhoods, would you be able to see that at the neighborhood level or would the size of the neighborhood mask the trend?

Jessie – On the 2nd, depends on your local context. Our neighborhoods have clear patterns within them. For the summaries it’s a great question. We have in mind currently to start with median rent per unit. Maybe like also average and mode and high and low. Or even number <unknown> to create a distribution. And then –So it’s actually interesting – we’ve done analysis with our current set which definitely includes duplicates to find the median at most scales. We found that the median is unaffected by deduplication. But we can’t report the number of listings accurately. That’s something we would hope to get to.

John K - The more granular may be the better. The neighborhood change is really about the edges where things are happening quickly, and that’s masked by any larger aggregation.

Katy – If you get it down to a small enough level, you probably have outliers.

Jessie – I don’t know that we’ve even talked about suppressing data. We would definitely report the number of listings in the aggregated area.

Eamon – So aggregation to the neighborhood or some kind of government area is helpful. But has anyone looked at trying to do clustering for neighborhood boundaries?

Jessie – Yes, for Boston’s use case it’s about certain initiatives. They might have a development that’s focused on low and middle-income housing. Which makes it harder to present summary data. You can always return to the land of the living and talk about these governed areas.

Katy – Are there other ways that you have been using the data?

Jessie – We have a ten percent threshold. If a municipality says the have 10 percent affordable, they get a particular program. But in order to get in, they have to have a plan but that plan includes needs for data and the like. So we incorporate our rental data into that. We incorporate that –

John K – Any quick reference to how that looks? Compared to Zillow?

Jessie – Compared to Costar – it’s a commercial property database…it’s almost like the opposite trends but we have questions about their data. Theirs is kind of a different market but I guess we would be interested in all of it. There’s large apartment buildings run by a property manager. It definitely…

Nancy – Have you started to scrape apartments.com?

Jessie – No, because we don’t want to get too many sites, because of deduplicates. But if there are sites that are focused on low income, those would be helpful.

John Cruz – Or some places where they use Zillow exclusively.

Jessie – I think we are getting better at permit data but it’s been a historical challenge.