Using Digital Signals To Measure Audience Brand Engagement At Major Sports Events: The 2015 MLB Season

Peter Ibarra, Dstillery
Peter E. Lenz, Dstillery

1. Introduction

Each year corporate brands budget huge sums of money for sports team sponsorships, but few methods exist to validate those expenditures. This research analyzes mobile device location data collected from all MLB stadiums during the 2015 season and the online browsing behavior associated with these devices. By marrying real world location data with online behavioral data, we are able to quantify engagement rates of in-stadium audiences and provide a measurement for the value of in-stadium advertising and team sponsorships.

To achieve this we:

1. Identify mobile devices for every baseball game by individual stadium.
2. Use a robust probabilistic matching algorithm to link mobile devices to a visitor's other devices, including desktop computers.
3. Calculate an affinity index for the users seen at each stadium, based on online engagement with a brand, interest, or market.

2. Data Collection

Data for our experiment originated from three primary sources: real-time bid requests from ad-monetized sites, online clickstream from third party application service providers and location behavior from software development kit (SDK) integrations. This unique combination of data is crucial to our being able to accurately sample online and mobile location behaviors. Using data from only one stream would greatly narrow our view of overall online behaviors and bring potential biases into our results.

Our native data collection system was originally developed for the collection of desktop only data. With the growth of smartphones, our system was extended to provide a probabilistic network of connections between all digital devices, also known as a device graph. With this we can tell what mobile devices are connected to what desktop devices, thus giving us a more robust view of the user's online behavior as they switch devices and locations throughout the day.

2.1. Real-Time Behavioral Data

We receive real-time online behavioral data from real-time bid requests and from third party desktop and mobile applications. Real time bid requests (BRQs) occur when a person is on an ad-monetized website or an app. A call comes from the site publisher or app publisher to fulfill an

advertisement slot. The call contains information such as an advertising device identifier (cookie, IDFA, AAID), a timestamp, an IP address, the publisher, the ad category and location. Not all fields are available in all instances of a BRQ. Third party desktop clickstream and mobile app clickstream are those data streams that are licensed from applications where users have opted-in to provide clickstream behavior in return for the functionality of the application. These data sets allow us to have a broader understanding of online behavior beyond ad-monetized sites and apps. [1] For convenience, we use the term "BRQ" to refer to records collected both through the RTB bidstream and through SDK integrations, as the type of information collected in each is similar.

2.2. Geolocation

Location, latitude/longitude, is one of the key signals collected from mobile apps either through the BRQs or third party licensing, For each MLB stadium in the study, we chose the center of the pitcher's mount to represent the centroid of the stadium and created a geofence around this centroid, translating a single point into a specifically crafted polygon. The geofence specifically contains the entirety of the baseball stadium grounds. Figure 1 below displays an example of the polygon methodology used in this experiment. A list of baseball stadium centroids (latitude and longitude) is included in appendix A.

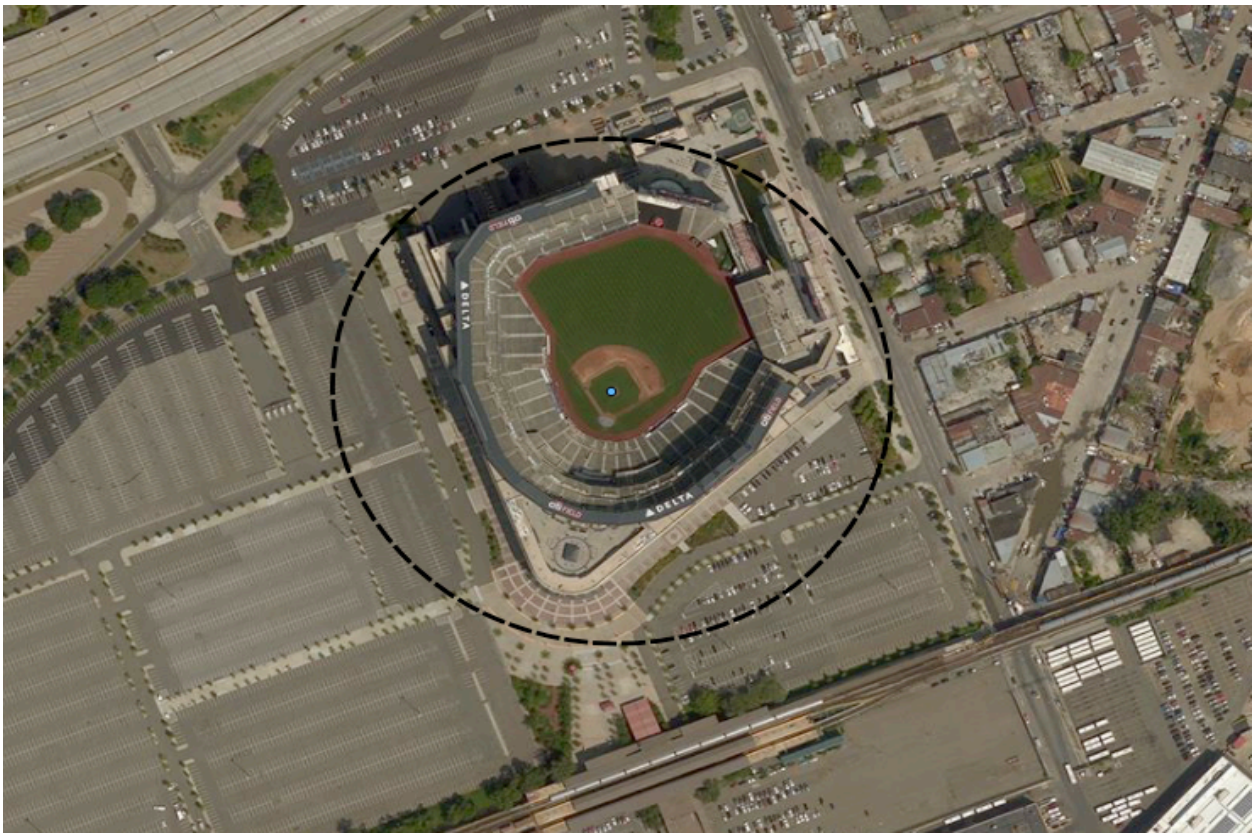

Figure 1. - Data collection geofence for Citifield

For each mobile RTB bid request and third party app clickstream that contained geospatial data during the 2015 MLB season, we matched location to our stadium geofence. If the BRQ location data fell into the geofence polygon dimensions, we qualified that data for inclusion in our experiment.

2.3 Device Graph

In order to get a holistic view of a user's online behavior, it is necessary to connect sample from the mobile data stream to the desktop data stream. We expanded our device location sample using a device graph. A device graph is a network of device IDs that probabilistically connects an individual user's devices to one another (smartphone, desktop, laptop, etc.). This bi-partite graph is built using a model that takes IP addresses, location data and time as inputs. Only wifi based IP signals are included in the graph. Devices that are seen on the same IP address within a specific period of time are considered connected. However, if too many devices are seen on one IP address than the system determines that the specified IP address is public access, a coffee shop or library for example, and prevents the IP from being used to expand a device graph. In this way we are able to avoid false positive connections in the device graph.

2.4 User Affinity

We quantify the online behavior of users in terms of user affinities. In this work, we deal with three types of user affinities: brand affinity, market affinity, and interest affinity.

Brand affinity for a given brand is defined simply by site visitation or app visitation to that brand's website or app, e.g. a user who has visited the brand's website is said to have a brand affinity for that brand. Market affinity and interest affinity describe whether the user is in-market for a given product or holds a certain interest, respectively, defined by certain online web or app actions. These are found using classification models, which determine whether each user has, or does not have each affinity. [2]

Our analysis frequently makes use of the index for a user affinity (either brand, market, or interest) for a given group of users with respect to the full population, as described in the results section.

2.5 Filtering Erroneous Data

To ensure the integrity of the BRQ and clickstream data collected, we run every BRQ event through a series of programmatic quality checks. An individual BRQ event can be polluted with good or bad faith errors. We applied three broad filters, Spatial, Behavioral, and Associative, to our data to eliminate any inaccurate or fraudulent data.

2.5.1 Spatial Filter

Many impressions carry spatial data that originates from a lookup based in IP address as opposed to a more accurate GPS calculated geo-location provided by a device. These IP-lookup based locations are by their very nature much less accurate than device-generated locations and usually locate to the centroid of the neighborhood, city or even state a device is located in. To combat this issue, our system counts the number of devices per unique location and calculates a distribution of these counts to discover and subsequently filter out these high-count IP-lookup locations. [3]

2.5.2 Behavioral Filter

The second filter we employed measures the distance and speed between locations of multiple BRQ or clickstream events from the same device. We filter out results that demonstrate impossible travel speeds between impressions. For instance, if an impression was seen in Austin, TX at 10:00 am and five minutes later in Miami, FL, our system detects this anomaly and all impressions from this device are filtered out.

2.5.3 Associative Filter

Our last quality check measures the integrity of the ad inventory being collected. Fraudulent ad impressions often occur on clusters of related websites. Theses sites - colloquially known as bot nets - generate revenue by showing ads to fully automated, computer-controlled, fraudulent users. Using machine-learning processes we are able to identify these clusters of fraudulent sites and the browsers and devices that visit them. Once these fraudulent devices are detected, we eliminate them from our dataset. [4]

2.6 Final Data Set

After our data has passed through these three filters, we are left with a dataset containing over 16 million unique and accurate location data points across the entire 2015 Major League Baseball season.

3. Experiment

3.1. Measurement Model

Our goal is to measure brand engagement rates for a Major League Baseball team's audience. To ensure the devices are reflective of attendees, we referenced the Major League Baseball schedule and collected devices only during scheduled baseball games. These measurements were carried out throughout the entire 2015 MLB season. The total set of users observed at each stadium is the "fan base" of the corresponding team.

For each set of data, we used our device graph technology to match the devices seen within each stadium to their home computers. Once the connection to home devices has been made, we have the ability to study each attendee's online browsing behavior and thus quantify the average browsing behavior of the "fan base" relative to every other MLB stadium and to the national average. The empirical probability of a website visit is determined by collecting two weeks of online behavior for every unique user seen at an MLB game. Each website visit for every unique URL is counted for the entire two week measurement period. Once the complete website visit count is calculated for the two week period, we determine the total visit count for a particular URL and divide it by the total population for the fan base. That is, the probability P of a visit to URL j for a team fan base i is given by:

$$P(j \mid i) = \frac{Z_{j,i}}{N_i}$$

where $Z_{j,i}$ is the sum of visits to URL *j* by users in fan base *i,* and N_i is the total number of users in fan base *i*. This probability can be interpreted as the probability of brand affinity for the brand corresponding to URL *j* for users in fan base *i*.

Similarly, we calculate *P(j)* for a random sampling of the full population, or the population of all MLB fans. We can then calculate an affinity index for this brand and fan base:

$$Ij,i = P(j \mid i) / P(j)$$

We use this index to measure brand sponsorship engagement rate in the sections that follow.

Analogous indexes can be calculated for the market affinity or interest affinity, based on the number of users in each fan base who show a given market or interest affinity according to our classification models.

3.2. Engagement Evaluation

To understand the impact of a corporate sponsorship on a team's fan base, we wanted to find sponsorships that were specific to a MLB team. For our experiment, we wanted to measure as accurately as possible the impact of a sponsorship by choosing sponsorships that were visible to an audience while attending the team's games. For this, we chose a major airline (sponsoring the Colorado Rockies), a national hotel chain (sponsoring the Chicago Cubs), and a national bank (sponsoring the New York Yankees). Each of these sponsors had in-stadium advertising, in-stadium seating sections and/or ticket promotions in conjunction with the team. Furthermore, these are brands that tend to lead to online browsing visits, enabling our brand engagement index measurement.

Once the sponsorships were chosen, we measured how each team's fan base interacted with the brands webpage. We calculated the brand engagement index, as described in section 3.1. This index quantifies how likely the team's fan base is to visit the brand's webpage compared to the population as a whole.

This method was broken out further by measuring the probability by month over the course of the season. That is, we considered the users seen in July, August, and September, as three separate sets, and applied the same methodology to calculate brand engagement index for each set. The general population's probabilities were also calculated separately for each month to ensure a consistent measurement, and brand engagement indexes were calculated for each month.

4. Results

4.1 Summary Statistics for MLB Audience

The baseball audience observed over the course of a season presented several unique characteristics. The number of devices seen at only one game at any stadium during the course of the season was between 20-30% across the entire MLB audience. Seventy percent of all devices were seen multiples times. The top three stadiums with the most unique audience were Dodger Stadium, Marlins Park and the Oakland Coliseum. The bottom three stadiums are Great American Ballpark, Kauffman Stadium and PNC Park. The five venues with the highest count of unique users were Dodger Stadium, Busch Stadium, Yankee Stadium, Target Field and PNC Park.

Two groups of fandom emerged within the MLB audience. The casual fan, which were users that were seen 1-2 times, and the avid fan, which were users that showed up at an MLB stadium 3+ times. Measuring the online browsing behaviors of the two groups against one another showed very distinctive characteristics.

The casual fan base visited college websites (.edu URLs) while also over indexing on entertainment and music content. This suggested an audience that was much younger in their demographics. While they were indeed attending games across the season, the casual fan browsing affinities suggest their stadium visits were more of a one-time entertainment option. On the other side, the avid group had online content suggestive of a much older audience. Their visitation rates over indexed on financial planning, vacation and investment content.

4.2 Singular Brand Affinity Score Over Time

For this experiment, we wanted to measure any change in an audience's brand affinity score over the course of the season. While scores varied over time, a general trend held true across brands that were also team sponsors. Brand affinity scores were higher among the MLB fan bases exposed to a sponsorship and generally increased as the season went on. A national hotel chain for the Chicago Cubs, a regional airline for the Colorado Rockies and a national bank for the New York Yankees were the three sponsorship brands studied for this experiment. Figure 2 shows the brand affinity scores for a stadium's audience for the months of July, August and September.

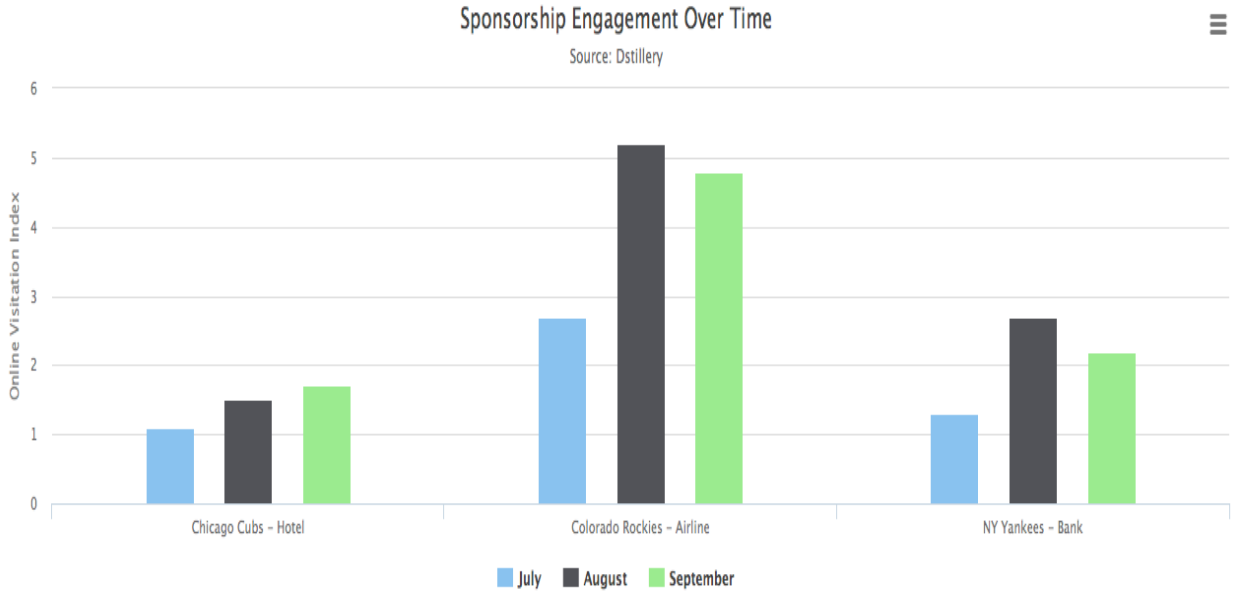

Figure 2. – Sponsorship Brand Affinity Score Over Time

The engagement rate was highest with the regional airline brand sponsoring the Colorado Rockies. The other two sponsorships did not see the same levels of affinity as the airline but still experienced growth as the season progressed. This growth in affinity scores could be attributed to prolonged exposure over the course the season. Additionally, the affinity scores suggest these audiences were not only becoming more aware of the brands but the audiences were also acting upon the exposure to the sponsorship as seen through their online behavior.

4.3 Per Stadium Audience Interest Affinity Scores

Beyond the measurement of a brand's affinity score, we were able to determine each audience's interest affinity, or the interests expressed through their online behavior. As shown in Appendix B, each audience's browsing interests could be indexed against the general population and a ranking, by venue, could be determined across all MLB stadiums.

These interests varied by region and team. Some of the affinity interest ranking were intuitive in their representation of the audience. The top five eco-conscious stadiums were Seattle, San Francisco, Oakland, Washington DC and Denver. The audience interests could also be used to measure the activities these groups participated. An example is the "Grill Masters" category that has a top five stadium list of Kansas City, St. Louis, Minneapolis, Houston and Dallas.

4.4 Per Stadium Audience In-Market Affinity Scores

Our analyses also extended to MLB stadium audience's differences in in-market or buying likelihood behaviors. . Using the same methodology employed in the previous affinity scoring, we measure each audience in-market to purchase product behavior as compared to the general population.

As shown in figure 3, these are the top five and bottom five audiences that are "In-Market" for cars.

Top and Bottom 5 Stadiums for "In-Market" Auto Audiences	
Automotive – Top Five	**Automotive – Bottom Five**
Pittsburgh	NY - Queens
Detroit	NY - Bronx
Tampa	Seattle
Atlanta	San Diego
Kansas City	Boston

Figure 3. – Top Five and Bottom Five In-Market Auto Stadiums

5. Conclusion

The merging of unique digital and physical datasets gives us invaluable insight into the relationship between teams, their fans, and their fans interests. Using our device graph technology in combination with geo-location data, we can understand where fans are browsing online and quantify how likely individuals are to engage with a particular brand. Additionally, using our affinity scoring methodology, each stadium's audience can be ranked by the interests of their fan base.

We built a model that provides a measurement system to be used by brand marketers and by Major League Baseball teams to understand the effectiveness of sponsorships. Measurements of brand affinity allow for analysis of previous sponsorships, as demonstrated here, and can also extend to inform future sponsorships, thus allowing brands to spend their marketing dollars more efficiently.

This methodology need not be limited to Major League Baseball. While this particular study was restricted to MLB stadiums, it can be expanded to all sporting venues. Similar analysis could be used to challenge or validate assumptions about which sport's fan base has the best prospective customers for a given brand. Additionally, this approach can be used to provide insight for individual teams or to gain more knowledge about a region (i.e. Los Angeles sports fans)

Finally, we envision applications beyond studies of brand sponsorship. We've shown that this model allows us to measure the interests of sub-groups of Major League Baseball fans, even within one team's fan base. Insights into the affinities and behavior of casual fans could inform the marketing strategies of the teams themselves in their efforts to increase game attendance.

We present a unique methodology that contributes to the understanding of how effective brand sponsorships are to a team's fan base, and can be extended to many more applications related to understanding the interests and behaviors of actual live game attendees. We believe our measurement system is valuable in shedding light into an area of advertising that has previously been difficult to quantify.

6. Acknowledgement

We would like to thank Dstillery and, most especially, the Dstillery data science team, for their invaluable feedback and insight. Their constant willingness to assist and answer any questions we had in our project allowed us to ensure the highest quality. We are deeply appreciative of all their contributions.

References

[1] Stitelman, Ori, et al. "Estimating the effect of online display advertising on browser conversion." Data Mining and Audience Intelligence for Advertising (ADKDD 2011) 8 (2011).

[2] Dalessandro, Brian, et al. "Scalable hands-free transfer learning for online advertising." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014.

[3] Provost, Foster J., Tina Eliassi-Rad, and Lauren S. Moores. "Methods, systems, and media for determining location information from real-time bid requests." U.S. Patent No. 9,179,264. 3 Nov. 2015.

[4] Stitelman, Ori M., et al. "Methods, systems and media for detecting non-intended traffic using co-visitation information." U.S. Patent No. 8,719,934. 6 May 2014.

Appendix A: Stadium Centroids

NAME	LATITUDE	LONGITUDE
Angel Stadium of Anaheim	33.800107	-117.883602
Globe Life Park in Arlington	32.751269	-97.082612
Turner Field	33.735395	-84.389544
Oriole Park at Camden Yards	39.283833	-76.621684
Fenway Park	42.346573	-71.097345
Wrigley Field	41.948267	-87.655445
U.S. Cellular Field	41.829347	-87.633788
Great American Ball Park	39.097239	-84.506537
Progressive Field	41.495984	-81.685285
Coors Field	39.756158	-104.994154
Comerica Park	42.339057	-83.048626
Minute Maid Park	29.75712	-95.355505
Kauffman Stadium	39.051579	-94.480345
Dodger Stadium	34.073849	-118.239951
Marlins Park	25.778053	-80.219427
Miller Park	43.028135	-87.971108
Target Field	44.981712	-93.277631
Citi Field	40.757039	-73.84588
Yankee Stadium	40.829615	-73.926351

O.co Coliseum	37.750891	-122.201576
Citizens Bank Park	39.906029	-75.166514
Chase Field	33.445486	-112.066682
PNC Park	40.446874	-80.005607
Petco Park	32.707533	-117.157057
AT&T Park	37.778397	-122.389341
Safeco Field	47.591445	-122.332366
Busch Stadium	38.622634	-90.192862
Tropicana Field	27.768311	-82.652324
Nationals Park	38.872971	-77.007459
Rogers Centre	43.641472	-79.389128

Appendix B: Per stadium affinity scores for select audiences

City	In-Market Auto	College Students	Eco-conscious	Foodies	Grill Masters	IT Decision Makers	Millennial Video Gamers	Federal Employees	Sports Fans-MLB
Atlanta	1.61543314	1.163312693	1.277065268	1.247835628	1.906976568	1.290513769	1.0797923	1.270952858	2.263969553
Kansas City	1.611742598	1.135410833	1.351425936	1.77876866	2.789742976	1.395889558	1.20914695	1.203480111	5.522899088
St Louis	1.585925283	1.182606388	1.362774888	1.71106895	2.289177945	1.316700319	1.249706757	1.177962643	4.585871911
Cleveland	1.539110721	1.638884833	1.185780523	1.528670245	1.720820975	1.354017521	1.334453951	1.058074335	3.458500989
Minneapolis	1.486585284	1.254110746	1.364294536	1.757243775	2.002705303	1.516414224	1.246840463	0.904452083	3.394556318
Anaheim	1.475390193	1.348199456	1.075532874	1.045222875	1.319797198	1.408686078	1.489277699	0.944683966	4.174919056
Cincinnati	1.453552801	1.838555063	1.059319472	1.418936603	1.702931672	1.244970512	1.318888525	1.044909235	3.561948125
Milwaukee	1.413118541	1.279274434	1.231350527	1.779707172	1.912442506	1.336358761	1.32664557	0.990154099	3.960408391
Denver	1.409370883	1.29371938	1.716195037	1.569795626	1.795831364	1.650824635	1.260535046	1.523270826	3.433208258
Arlington	1.391553455	1.214653262	1.093511679	1.120063191	1.942465898	1.356713385	1.243248229	1.113495752	2.985000392
Philadelphia	1.37446675	1.62172844	1.152320993	1.452344994	1.558068743	1.321786433	1.283019623	1.181721672	4.465407928
Chicago	1.36228612	1.597789532	1.243895881	1.454219893	1.901221573	1.580911898	1.237251161	1.158718313	4.404645152
Houston	1.347733632	1.309092933	1.064175606	1.216249273	1.96737896	1.418757615	1.367411114	0.888664532	3.012802954
Phoenix	1.313994067	1.059720315	0.999335902	1.189191721	1.468511292	1.264890417	1.301192944	1.124585021	3.50855192
Oakland	1.305153809	1.818479031	1.953735962	1.064515591	1.507595141	2.014373326	1.383209046	1.143321763	4.738102945
Washington DC	1.289511352	1.854315071	1.827771896	1.487474945	1.770528295	1.863790972	1.176072569	3.929848045	4.76397186
Miami	1.284127062	1.259635576	0.837502385	0.933144751	1.205332436	1.445526436	1.296370518	0.983497888	3.310309619
San Francisco	1.275610256	1.691293868	1.955268928	1.182841102	1.473357187	2.232116756	1.41117386	1.245145218	4.179091469
Los Angeles	1.249024137	1.494358562	1.174339856	0.842469056	1.020896214	1.437928833	1.34002626	0.805676775	3.887593806
Baltimore	1.23943665	1.688584321	1.492005986	1.547006376	1.764767059	1.571995732	1.420764354	2.697981063	5.196077655
Chicago	1.196212059	1.53952514	1.236020839	1.345080388	1.745182242	1.428429017	1.351458454	0.907698868	4.144872678
Boston	1.177468604	1.830570896	1.421964944	1.418949876	1.386620545	1.647508757	1.209790951	1.142092367	3.210199449
San Diego	1.17692233	1.374564621	1.247138254	1.070863625	1.402523525	1.665761672	1.457297593	1.291021172	4.54669022
Seattle	1.172163638	1.392839922	2.266116819	1.555901479	1.652078907	2.039383545	1.603368615	1.245575484	3.093208891
New York/Bronx	1.158837674	1.963662757	0.985894701	1.138681426	1.162617542	1.369219259	1.304374907	0.986139032	3.155122085
New York/Queens	0.997411854	2.164007749	0.992494445	1.110289196	1.132515199	1.437053053	1.313750356	0.910424027	4.104032153
MEDIAN	1.330863849	1.443599242	1.23995836	1.382008495	1.711876324	1.432741035	1.309062632	1.133338694	3.924001098
MAX	1.61543314	2.164007749	2.266116819	1.779707172	2.789742976	2.232116756	1.603368615	3.929848045	5.522899088
MIN	0.997411854	1.059720315	0.837502385	0.842469056	1.020896214	1.244970512	1.0797923	0.805676775	2.263969553