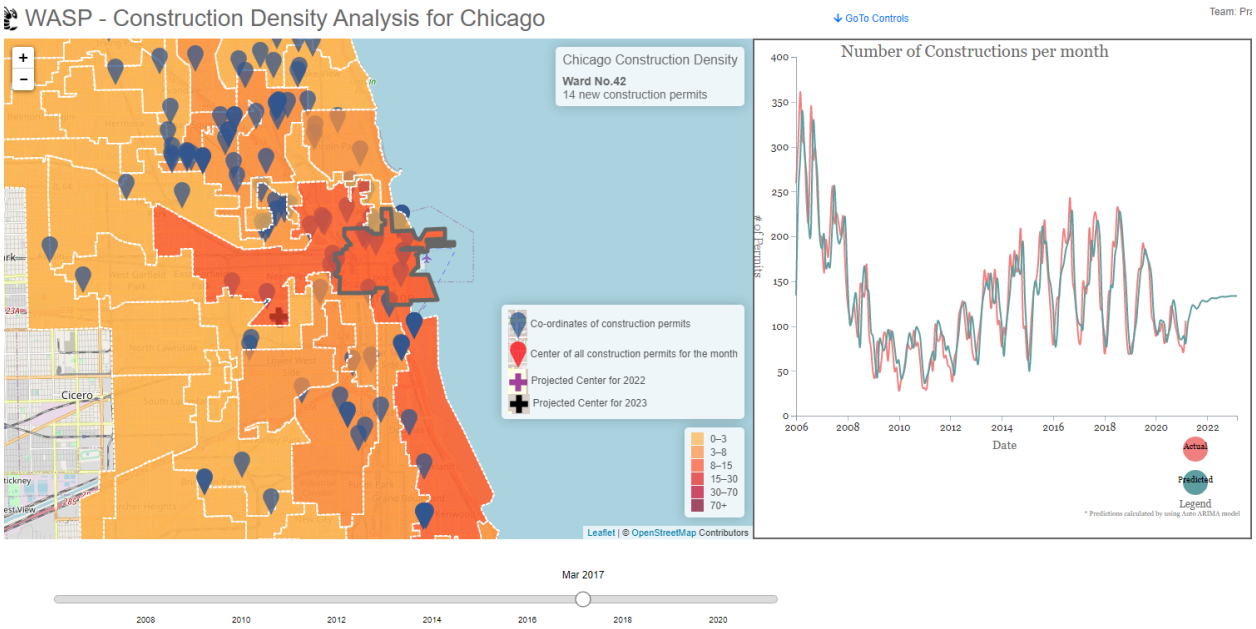
WASP - Construction Density Analysis for Chicago

Aayush Parwal, Brian Tran, David Scott Bader, Prashant Kubsad, Wael Ahmad Sultan



Project website: https://dva-wasp.s3.amazonaws.com/workspace/index2.html

Introduction - Motivation

Our aim was to build an interactive portal to visualize construction density trends in an urban area and perform time series analysis. Currently, there are many static visualization websites based on either geo-spatial or temporal, but not both. Existing analyses focus on prices mainly, rather than density of constructions. Real Estate Analysts can project growth based on density movement. City planners can make data driven decisions for new infrastructure development. Our tool was designed to reduce the decision making time of stakeholders by aggregating construction trends and time-based analysis.

Problem Definition

Currently there are no free tools or services that combine geo-spatial and temporal analysis of construction data. Such an analysis, extended to factors beyond pricing, like density can greatly reduce decision making time and add value to our users.

Survey

A spatio-temporal analysis done in the article [PK-1] provided several parallels in our aim to visualize the construction trends over time. Much work has been done with choropleth maps to visualize geo spatial models, like dynamic increase in perceivable area [PK-2], boundary neighbor selection [PK-2]. Coupled with Google Maps/API, gives the capability to develop interactive web pages. Reactive time component to geo-spatial models presents its own challenges. Possible solutions are discussed in EST [PK-3].

First two papers gave us guidelines on ML models to use to predict growth such as ARIMA, exponential smoothing [W-1] and mix of Markov chain and Cellular Automata [W-2]. Last paper helped guide on visualization types and their best use [W-3]. To overcome potential complexity challenges for both visualization and modeling we use third party platforms as a service.

One study [AP-1] addressed how construction permits for residential, commercial or public buildings correlate with socio-economic demography of an area. Study cites major challenges to read, manipulate and store large amounts of detailed data required for any geo-spatial analysis. With cloud computing, we reduce such limitations. The study [AP-2] identifies damage and recovery efforts based on building permits and spatial scans. Our tool could enable city planners to balance approving building permits by understanding clusters and allocate resources accordingly. Another study [AP-3] used density of population to dynamically adjust k value in the algorithm within city concentration of building permits as needs change.

Research using construction data yielded meaningful insights on trends and event linkage [SB-1] [SB-5], that we can build on. Some earlier efforts used outdated technology (e.g., ESRI ArcGIS) and outdated methods (e.g., MSExcel) to organize data [SB-4]. Other efforts used effective data analytics techniques, but deficient visualizations [SB-2]. We can improve visualization by replacing static diagrams with interactivity and better practices [SB-2].

[BT-1] Researchers proposed forecasting with construction terms from Google Trends. Our forecasting model is subject to data lag and we could supplement our forecasting model with search terms similarly. [BT-2] Bagshaw compares 4 forecasting models. In our project, we used a TimeSeries forecasting model. This paper serves as a foray into several popular models. This paper [BT-3] proposed a methodology for assessing community health based on infrastructural investment. The researchers established data processing conventions we can adopt on our data set. The researchers failed to establish a causal relationship.

Proposed Method

Intuition

Based on our combined research and subsequent discussion, we proposed the following:

1. Analytics based on density of constructions using time series analysis, as opposed to existing analyses that focus on prices and are factor based.
2. Introduce the novel idea of merging geo-spatial with temporal analysis.
3. Provide auto play and pause capability to the user on the data timeline enabling interactivity to visualize construction densities over time and space.
4. Show construction density trends using mean latitude and longitude data.

Approach

Data: We started with construction permit data for the city of Chicago. The city's website has construction permit data starting in 2006 with more than 657,000 rows across 119 columns.

Custom code was written in Python to clean and extract the required data set in csv format and GeoJSON format. The full data set contained information about permits that were not critical to our analysis, like renovation, and electric working. We decided to consider only new construction permits for our analysis. Data is fetched in real time with the user's monthly period selection embedded in the API "get" query. This gave us the advantage of working on smaller data size for the visualization and time series analysis.

Visualization: We opted to use D3.js as our primary visualization library along with Leaflet (an open source map library). Leaflet tiles are easy to use and provide us with many layers of out-of-the-box visualization elements, e.g., street names, roads, and state boundaries. We enhanced this basic tile to include polygons representing each Ward in Chicago. D3 combined with the Leaflet.js library provided an exceptionally strong platform to visualize spatial components. For the temporal component, we used a D3-based slider, where each tick represents a month from January 2006 to March 2021. When the cursor stops at a particular location, all the new permits for that month are filtered from clean data. Latitude and longitude for each permit is transformed onto the map. There is a transition element in the slider that slowly moves over the timeline and animates how construction permits have grown over time. We built our interactive visualization as a web application and hosted on AWS S3 to render the site on a browser. Leaflet and Bootstrap makes the website compatible on mobile devices.

Algorithms: We performed two types of time series analysis on the data

1. Vector autoregression forecasting of construction hotspots

One facet of our construction forecasting revolves around making future predictions solely from geospatial data. Contemporary approaches at construction forecasting rely on factor based modeling such as decision trees trained on median income, proximity to water, and infrastructure investment. We utilize a vector autoregressive model (VAR) constructed with latitude and longitude data of construction permits in order to predict the location of the construction hotspots up to two months in the future. We do this by calculating the mean coordinate data for all construction, grouped by year to obtain the location of the “hotspot” as it moves over time. This data serves as the training set for a VAR(p) model which models subsequent hotspot locations as a linear function of previous coordinates up to and including pth order lags:

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + e_t,$$

The above equation is for VAR(p=1). We determine the optimal value of p by exhaustive search and selecting the order lag which corresponds to the highest Akaike information criterion (AIC). Once the model is tuned with the optimal p, we query for the hotspot coordinates for future years beginning with 2022.

2. ARIMA for Construction Permit Forecasting

With the intention to understand the trend of new construction permits in the city of Chicago, we performed time series forecasting using the ARIMA model on the data.

We first removed unnecessary columns from the source data which was read in as csv but date columns parsed as datetime with the dataframe indexed with the datetime column as well.
We then aggregated our data, grouping it month-over-month from 2006 to March of 2021.
We used auto arima function to fine tune hyper parameters (p, d and q) for the model.

Experiment(s) / Evaluation:

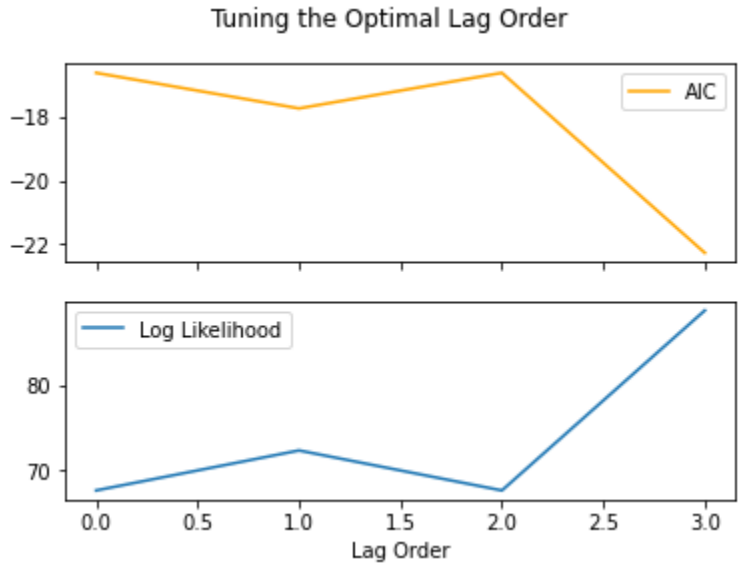
For the Interactive Visualization we iterated our design, trying out various ideas and several different layouts, rejecting some, keeping others with refinements.

We evaluated D3's choropleth, Google MAP api and Leaflet open tiles for the base map. After testing the integration between D3, JQuery and map, we selected Leaflet because it was easier to add layers on the map for density, street information,and combine it well with D3 markers and sliders. Leaflet also makes the maps reactive to mobile and smaller devices.

We enhanced the map by including a heatmap view based on polygons representing each ward in Chicago and its respective number of building permits. Users can switch from heatmap visualization to more detailed view with pins representing building permit locations. For the temporal component, we included a slider to navigate between monthly data. We added a button to animate how construction permits have grown over time including the forecasted period. On the right side we show the building permits trends and an option to get a couple of years forecast.

For both the VAR Model and ARIMA Model we used the auto tuning function in the respective library to find the best parameters for our models. We relied on MSE and AIC respectively to validate the optimum parameters selection and eventually the prediction performance.

VAR Model: Our experimentation with the VAR model aims to achieve three things. First, to identify the optimal hyperparameter p, which determines the order of lags to use in autoregression. Second, to assess the quality of the model against a holdout data set. Finally, to forecast hotspot locations for the years 2022 and 2023.



1. We tuned hyperparameter p using an exhaustive search methodology. Our search space was all integer values from p=0 to p=3 (inclusive). For each value of p, we trained a VAR model and calculated the AIC and log-likelihood (LL). Ultimately, we used the p value associated with the lowest AIC which in this case, was p=3.

2. With the optimal hyperparameter p=3 identified, the next step was to build a model from the training data (years 2006-2017) and make predictions for the holdout set, which in this case were the coordinate data for years: 2018, 2019, 2020, and 2021. We quantified model performance by the geographic distance between the predicted location and the actual location shown in the far right column below.

	Actual		Predict		
Year	Lat	Long	Lat	Long	Distance (mi)
2018	41.8505	-87.6624	41.8857	-87.6683	2.6
2019	41.8467	-87.664	41.8697	-87.6795	2.4
2020	41.8702	-87.6678	41.8645	-87.6805	1.0
2021	41.8663	-87.6703	41.8714	-87.6781	0.8

3. Our final model was built using all the data (2006-2021) with the order lags set to p=3. With this model in hand, we made forecasts for the next two time steps, yielding the following projections for 2022 and 2023:

Year	Latitude	Longitude
2022	41.872106	-87.666882
2023	41.867022	-87.675163

Observations

1. **Interactive Visualization:** If the map is zoomed out way too much, the markers get concentrated to small areas and do not give valuable information. That's where the option of switching between density map vs markers is very useful to turn off noisy markers. Leaflet library's integration with D3, Bootstrap and JQuery was seamless. Since the markers were driven by API calls, we had to calibrate the speed at which the timer moves to make sure we get response for each month and visualization is updated before moving onto the next date.

2. **ARIMA Model**: We split our data into training and test, and trained the ARIMA model with the best parameters from the previous step based on AIC scores. We then trained the model on all of the data and can now forecast new future permits. It's important to note that for forecasting, we had to create an index time-range into the future and then use that range as the index. It's also important to note, we are not just able to forecast future permits, but also plot model expectations with existing data.

3. **VAR Model:**

Hyperparameter candidate selection: Because the data source began in 2006, and this model operated on aggregated annual data, we had only 14 data points with which to train and test the VAR model. This dearth of data proved especially problematic during hyperparameter tuning as the lag order p is non-trivially limited by small data sets such as this. As a result, the

search space for p was limited to p <= 3. With more data, it is possible that the optimal value of p is greater than 3.

Limited forecast period: We originally intended to provide accurate forecasts many years into the future (extending into the 2030's). However, the VAR model faced the same limitation as many other autoregressive schemes, which is that only the first few time steps out provide useful predictions. In our analysis, we found that the predictions for 2022 and 2023 were useful. Past that, the predictions were identical for every single year, which is why we exclude these predictions from our final results.

Quality of model: Evaluation of the final model on years 2018-2021 yielded errors (measured geographically) within the range [0.8, 2.6] miles. The 95% confidence interval on the error is 1.7 ± 0.8 miles. We assess the severity of these errors by evaluating them in the context of Chicago's land footprint. Chicago's largest north/south diameter is 25 miles and its largest east/west diameter is 15 miles across. Together, these two measurements comprise its area of 228 sq. miles. Generally speaking, we find that these errors are acceptably small in the context of Chicago's overall dimensions. However, we note that Chicago's urban density may magnify the weight of the errors as the span of 1.7 miles could cover an area with several hundred buildings and thousands of people.

Conclusion and Discussion

We achieved our original objectives in building out a minimal viable product (MVP).

1. Existing analyses are mainly focused on prices and are factor based whereas our approach is based on density of constructions using time series analysis.
2. Novel idea of merging geo-spatial with temporal analysis.
3. Provide auto play and pause capability to the user on the data timeline enabling interactivity to visualize construction densities over time and space.
4. Construction density trends using mean latitude and longitude data.

Innovation: Our tool (called WASP) is a spatio-temporal visualizer that shows construction density trends based on analyzing cities' building permits data, in our case the city of Chicago. The tool also performs time series analysis to forecast not only the number of monthly permits but also the anticipated hot spots, another differentiating from other websites we observed. Tool is beneficial for a wide range of users including real estate analysts, construction developers and municipalities. Users can draw insights on density movements, and lucrative development areas which could help reduce the decision making time and save the hassle of jumping from one tool to another.

Future Directions: Ample opportunities exist to expand on our work, such as extending to include other cities and perform benchmarking analysis.

Distribution of team member effort

All team members have contributed a similar amount of effort.

References

[PK-1]	Using Building Permits to Monitor Disaster Recovery: A Spatio-Temporal Case Study of Coastal Mississippi Following Hurricane Katrina https://www.tandfonline.com/doi/abs/10.1559/152304010790588052
[PK-2]	PK-2: Dynamic Choropleth Maps – Using Amalgamation to Increase Area Perceivability https://ieeexplore.ieee.org/abstract/document/8564174
[PK-3]	Exploratory spatio-temporal visualization: an analytical review Journal of Visual Languages & Computing, Volume 14, Issue 6, December 2003, Pages 503-541 https://www.sciencedirect.com/science/article/pii/S1045926X03000466
[WS-1]	Smart transportation planning: Data, models, and algorithms https://www.sciencedirect.com/science/article/pii/S2666691X20300142
[WS-2]	HomeSeeker/ A visual analytics system of real estate data https://www.sciencedirect.com/science/article/pii/S1045926X17301246
[WS-3]	Spatiotemporal urbanization processes in the megacity of Mumbai, India: A Markov chains-cellular automata urban growth model https://www.sciencedirect.com/science/article/pii/S0143622813000362
[AP-1]	The Future of Spatial Analysis in the Social Sciences https://www.tandfonline.com/doi/abs/10.1080/10824009909480516
[AP-2]	Using Building Permits to Monitor Disaster Recovery: A Spatio-Temporal Case Study of Coastal Mississippi Following Hurricane Katrina https://www.tandfonline.com/doi/abs/10.1559/152304010790588052
[AP-3]	Adaptive clustering algorithm based on kNN and density https://www.sciencedirect.com/science/article/pii/S0167865518300266
[SB-1]	Rubén Hernández-Murillo, Michael T. Owyang, and Margarita Rubio. 2017. Clustered housing cycles. *Reg. Sci. Urban Econ.* 66, (2017), 185–197.

[SB-2]	Massimo Cecchini, Ilaria Zambon, and Luca Salvati. 2019. Housing and the city: A spatial analysis of residential building activity and the Socio-demographic background in a Mediterranean city, 1990–2017. *Sustainability* 11, 2 (2019), 375.
[SB-4]	Melissa Shakro. 2013. Tracking neighborhood development and behavioral trends with building permits in Austin, Texas. *J. Maps* 9, 2 (2013), 189–197.
[SB-5]	Margherita Carlucci, Efstathios Grigoriadis, Giuseppe Venanzoni, and Luca Salvati. 2018. Crisis-driven changes in construction patterns: evidence from building permits in a Mediterranean city. *Hous. Stud.* 33, 8 (2018), 1151–1174.
[BT-1]	Now-Casting Building Permits with Google Trends Coble, David and Pincheira, Pablo M., Now-Casting Building Permits with Google Trends (February 1, 2017). Available at SSRN: https://ssrn.com/abstract=2910165 or http://dx.doi.org/10.2139/ssrn.2910165
[BT-2]	Univariate and Multivariate Arima Versus Vector Autoregression Forecasting Bagshaw, Michael L., 1987. “Comparison of Univariate ARIMA, Multivariate ARIMA and Vector Autoregression Forecasting,” Federal Reserve Bank of Cleveland, Working Paper no. 86-02.
[BT-3]	The Other Side of the Broken Window: A Methodology that Translates Building Permits into an Ecometric of Investment by Community Members O’Brien, D.T., Montgomery, B.W. The Other Side of the Broken Window: A Methodology that Translates Building Permits into an Ecometric of Investment by Community Members. Am J Community Psychol 55, 25–36 (2015).