Introduction - Motivation

Our aim is to build an interactive portal to visualize construction density trends in an urban area and perform time series analysis.

Currently, there are a lot of static visualization websites based on either geo-spatial or temporal but not both. Existing analysis are mainly focused on prices than density of constructions.

Real Estate Analysts can project the growth based on density movement. City planners can make data driven decisions for new infrastructure development. This tool will reduce 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 will greatly reduce decision making time and add value to our users.

Survey

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

The first two papers provided us guidelines on what ML models to use to predict growth such as ARIMA, exponential smoothing [W-1] and mix of Markov chain and the Cellular Automata [W-2]. The last paper was more focused on types of visualization and their best of use [W-3]. Lastly, to overcome associated potential challenges related to complexity of both the visualization and modeling we plan to use third party platforms as a service.

The study[AP-1] mainly wants to address how construction permits for residential, commercial or public buildings correlate with socio-economic demography of an area.Study cites major challenges in being able to read, manipulate and store large amounts of detailed data which is required for any geo-spatial analysis.With flexibility and cloud computing, we reduce such limitations.The study[AP-2] identifies damage and recovery efforts based on building permits and spatial scans.Our tool aims to enable city planners to balance giving out building permits by understanding disaster recovery clusters and allocate resources accordingly.The study[AP-3] looks to utilize density of population to dynamically adjust k value in the algorithm as even within city concentration of building permits needs to be changed for example, city center vs suburbs etc.

Previous research using construction data, shows that it can yield meaningful insights, in terms of trends and event linkage[SB-1][SB-5]. We seek to build upon earlier efforts. 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 (avoid red-green color schemes. [SB-2].

[BT-1]The Researchers propose 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 will use a TimeSeries forecasting model. This paper serves as a foray into several popular models. This paper[BT-3]proposes a methodology for assessing community health based on infrastructural investment. The researchers establish data processing conventions we could on our data set. The researchers fail to establish a causal relationship.

Methodology

Data:

We started looking for construction permit data for the city of Chicago. The city's website had all the construction permit data going all the way back to 2006. This data has more than 657,000 rows across 119 columns.
Custom code is written in Python to clean and extract the required data set in csv format and GeoJson format. The data set contained information about permits that were not critical to our analysis, like, Renovation, Electric Working. We decided to consider only New Construction permit type for our analysis.

Visualization:

As demonstrated in home-work2, we opted to use D3.js as our primary visualization library along with Leaflet (opens source map library).
Leaflet tiles are easy to use and provides us with many layers of visualization out of the box, ex.: street names, roads, state boundaries. We have further enhanced this basic tile, to include polygons representing each zip code in chicago.D3 combined with Leaflet.js library provides us with a very strong platform to visualize spatial components.
For the temporal component, we are using a D3-based slider, where each tick represents a month between Jan 2006 to March 2021. When the cursor is stopped 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. A web application is built for this and is hosted on AWS S3 to render the site on a browser.

Analysis:

We are performing 2 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 month (beginning in 2011) 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). Our VAR model currently uses a pth order lag proportional to the number of observations but this may be adjusted in future model tuning.

2. **ARIMA for Construction Permit Forecasting**

With the intention to understand the trend of new construction permits in the city of Chicago, we looked to perform 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.

We then 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 into the future, 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.

<u>Upcoming Experiment(s) / Evaluation:</u>

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. We are exploring options to show markers for each

point when the map is zoomed to a certain level, else show a heat map for each zip code showing the number of constructions.

2. **VAR Model**

   Parameter tuning and model evaluation are ongoing, but we have some promising preliminary results regarding hotspot location forecasting.
   To avoid look-back bias in our model, we used the first 80% of the data as the training set and the last 20% (up to March 2021) for validation. For our first slate of tests, we chose the default pth order lag from the python “statsmodels” library to fit the model VAR(p=12). After training, we calculated the mean squared error as well as the worst error in terms of geographical distance for the one-month and two-month projections.

Conclusion and Discussion (*under progress*)

Plan of Activities

Activity	Assigned	Date - OLD Plan	Date - Revised Plan
Data Gathering, Clean Up and Storage	All	03/26/2021	03/26/2021 ***Completed***
POC for visualization with fixed data	Pkubsad3, wsultan3	04/09/2021	04/01/2021 ***Completed***
Visualization with time scale and data filtering	Pkubsad3, wsultan3	-	04/10/2021
Time series analysis of data (ARIMA)	aparwal7	04/01/2021	04/01/2021 ***Completed***
Time series analysis of data (VAR)	Btran411	04/15/2021	04/15/2021
Final product with analysis	All	04/25/2021	04/25/2021
Poster Presentation and Final Report	All	05/01/2021	05/01/2021

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).