Using Geolocated Text to Quantify Location in Real Estate Appraisal

Tim Niklas Heuwinkel
Paderborn University (UPB)
tim@heuwinkel.de

Jan-Peter Kucklick
Paderborn University (UPB)
jan.kucklick@upb.de

Oliver Müller
Paderborn University (UPB)
oliver.mueller@upb.de

Abstract

Accurate real estate appraisal is essential in decision making processes of financial institutions, governments, and trending real estate platforms like Zillow. One of the most important factors of a property's value is its location. However, creating accurate quantifications of location remains a challenge. While traditional approaches rely on Geographical Information Systems (GIS), recently unstructured data in form of images was incorporated in the appraisal process, but text data remains an untapped reservoir. Our study shows that using text data in form of geolocated Wikipedia articles can increase predictive performance over traditional GIS-based methods by 8.2% in spatial out-of-sample validation. A framework to automatically extract geographically weighted vector representations for text is established and used alongside traditional structural housing features to make predictions and to uncover local patterns on sale price for real estate transactions between 2015 and 2020 in Allegheny County, Pennsylvania.

1. Introduction

Many different stakeholders depend on real estate appraisals to support their decisions or to calculate revenues and cost, which naturally creates a great need for accurate valuation of property. For most individuals, buying a house is one of the greatest financial decisions in their life [1, 2]. Governments around the world tax property based on value [3], for example making up 12.1% of all tax revenue in the United States of America [4]. Banks also calculate credit cost and mortgage payments based on private property value [5]. Recently emerged trading and renting platforms for houses and apartments like Zillow, Realtor or Airbnb make use of automated real estate appraisal to help customers find fair prices and reduce information asymmetry [6]. Lastly there is a massive market for financial investments in real estate. Institutional portfolios allocate, on average, 5.1% of their total value in real estate [7]. Better and automated evaluation models powered by Machine Learning (ML) can be used to easily check prices of properties to buy, decrease costs for calculating taxes, gain an edge over competitors or verify real estate assessments.

Aside from physical attributes of a house like size or age, one of the largest aspects making up the sale price of a property is its physical location [2, 8, 9, 10, 11]. There are mainly two approaches to incorporate location data into real estate valuation: model-driven and feature-driven approaches.

Model-driven approaches create sub models for different areas and make use of techniques like spatial regression and Geographically Weighted Regression (GWR) [12]. Meanwhile feature-driven approaches try to capture as much information about the environment as possible [9] by creating more sophisticated features like the spatial distance to certain Points Of Interest (POI) or the amount of POI inside a buffer around the house. POIs can relate to businesses, i.a. bars and grocery stores, as well as to education, i.a. schools, or environmental aspects like green spaces or water bodies [9, 11, 13]. One recent shift within the feature-driven models is to include unstructured data like satellite or street-side images, for which the location information is often extracted by Convolutional Neural Networks (CNNs) [6, 14, 15, 16].

Since 80% of data exists in an unstructured form, much of which in form of text [17], text should also be considered as a data source for real estate appraisal. While some previous authors [2, 15] made use of textual descriptions of the house's attributes, to the best of our knowledge, no work exists examining the location based on text analysis. Inspired by the work of Sheehan et al. [18], who use geolocated Wikipedia articles for poverty prediction, we assess whether they can enhance quantifications of a property's location to improve predictive performance of ML models for real estate appraisal. Geolocated Wikipedia articles are

URI: https://hdl.handle.net/10125/80039
978-0-9981331-5-7

about real life objects or locations, like universities, parks, train stations, bridges etc. that have received a user curated geotag, representing the global position of said object or location.

Our results indicate that incorporating text data from Wikipedia articles into ML models improves predictive performance in Spatial Out-Of-Sample (SOOS) validation by 8.2% over a traditional GIS baseline. The most important spatial features in our models relate to accessibility, private businesses, recreation and education.

Our contribution is three-fold: First, we contribute to the existing literature by applying insights from text regression to real estate appraisal. Second, we propose a framework to process and analyze geo-tagged text data, which includes a novel modification of the term-frequency inverse document frequency (TF-IDF) formula, which makes it possible to model spatial dependencies. Third, we show the efficacy of Wikipedia articles as a source of geotagged data for real estate appraisal.

The remainder of this paper is structured as follows: The next section provides background on the topic of real estate appraisal, geospatial analytics and ML. Section three introduces the dataset, the proposed framework, modeling process, and the empirical results. Section four contains a discussion of the results in geospatial context and compares findings with the existing body of literature. We conclude by pointing out limitations, as well as implications and offering directions for future research.

2. Related Work

In this section, we shortly summarize the related work about real estate appraisal, geospatial analytics for real estate appraisal and the foundations of ML and Natural Language Processing (NLP).

2.1. Real Estate Appraisal

Most methods for automated real estate appraisal are based on the hedonic pricing theory [19, 20]. In real estate appraisal, hedonic pricing theory assumes that the value of a property can be represented as an aggregation of discrete characteristics (Linear Regression). The original proposal of Lancaster [19] is that the overall utility of a commodity is measured by the utility of its characteristics. Rosen [20] then built upon this concept and suggested that prices reflect the demand and quality of those characteristics. One desiderata of real estate appraisal models is interpretability, since many stakeholders rely on these models to gain insights into the domain or to justify price estimates to their customers [14]. The price utility estimation of distinct characteristics in the hedonic model satisfies this need for interpretability.

Typical non location-based structural features in real estate include, but are not limited to: lot size, age, condition, room count, floor area on certain floors, garage space, type of house or building style and amenities like fireplaces, heating and air conditioning, pools etc. [1, 9, 5, 21, 22].

Aside from the structural characteristics of a house, the property's location is an essential characteristic, making up a large portion of the sale price and the value [2, 8, 9, 10, 11]. Although it is important, location is hard to integrate into ML models. The underlying problem is the lack of a simple quantification. For example, housing attributes like size, age or rooms are simple to quantify in units like square feet, year or count, while the value of location is more complex. In contrast to simple spatial membership (e.g. dummy coded neighborhoods), creating feature vectors descriptive of the environment (e.g. distance to POI) can enable more fine granular comparisons.

2.2. Geospatial Analytics

Geographical Information Systems (GIS) are systems designed to save, edit, analyze and visualize geographical data and can help to incorporate spatial data in real estate appraisal. According to Anselin [8] researchers and practitioners in real estate appraisal agreed that real estate markets have a spatial nature which was not reflected by the research efforts at the time and that there is a need for more practical frameworks to incorporate spatial data. The first law of geography is ”everything is related to everything else, but near things are more related than distant things.” according to Tobler [23]. This phenomenon is also called ”spatial autocorrelation” [24, 25] and is one of the main pillars of GIS applications in real estate appraisal, since the value of location of two houses are similar, if they are close to one another. This poses challenges and possibilities. On one hand, many standard statistical testing and analysis processes assume independent variables, which is not the case in GIS applications due to spatial autocorrelation. On the other hand it is a partially predictable phenomenon, which can be exploited to improve existing models [25].

Model-driven approaches try to incorporate spatial autocorrelation into real estate appraisal by using a sophisticated modeling procedure. One of the first approaches was proposed by Anselin [26] and is called spatial regression. In spatial regression, a spatial weight matrix, containing weights of the neighboring areas, is used to reduce autocorrelation. However, spatial dependencies are not the same for every distribution. Thus, Brunsdon et al. [12] improved the model by accounting for spatial heterogeneity with a method he called Geographically Weighted Regression (GWR). GWR captures spatial heterogeneity by estimating coefficients of local models with a locally weighted average.

Another popular technique with GIS is to analyze spatial fixed effects [27]. Spatial fixed effects can be analyzed by creating individual models for each subgroup or distinct areas in the data. They can also be incorporated into a single regression model by simply adding a binary variable for each spatial group, a process also called dummy coding [5]. This approach can be quite effective in improving accuracy for models in real estate appraisal [28, 29, 30]. Simply incorporating spatial fixed effects may even eliminate the need for more advanced model-driven approaches like GWR in some cases [10, 11].

While modeling spatial fixed effects may suffice for model-driven approaches, feature-driven approaches make use of more sophisticated feature representations of the general environment, which can increase predictive performance [2, 9, 13, 29, 31]. One possible representation is to include the spatial distance to the house or the amount of POIs from each category around the house. The distance can be measured in units of length or time to reach the nearest POI of each category. The categories can be grouped into four meta-categories: accessibility (i.a. public transport, highway, airport) [2, 13, 29, 31], recreation (i.a. parks, nature, bars, music venues) [2, 9, 13], education (i.a. schools and higher education) [2, 9, 13] and private businesses (i.a. shops and services) [2, 9].

Yet, calculating distances to POIs from GIS databases is not the only way to represent the environment. Sheehan et al. [18] have shown that geolocated text in form of Wikipedia articles can be used to create vector representations of articles which were successful in predicting economic development.

An advantage of geolocated Wikipedia articles as a data source is that, unlike GIS databases, which usually only contain information about the existence and the position of an object, Wikipedia articles also offer rich contextual information in form of text for each object. It can therefore be hypothesized that those articles can be useful for characterizing environments in real estate appraisal as well.

Another advantage of text data is availability. A major portion of data on the web is in form of text [17] and can currently be seen as an untapped reservoir for real estate appraisal. Availability of high quality GIS data is not guaranteed for every region [18]. Even though it is increasing due to Open Data Portals for city data, the coverage is still sparse, or hidden behind a paywall. This is where text data can help to fill gaps in coverage or to improve performance by supplying additional information.

2.3. Machine Learning

While text data is often widely available, representing it in a way that is understandable by ML algorithms requires some preprocessing. The first step in many NLP pipelines is tokenizing, which means breaking up sentences or documents into a list of single words. After creating the tokens, n-grams can be created, which are sequences consisting of n words, which appear often together and are sometimes more informative than single words [32]. To obtain usable feature sets from text data, the text from the Wikipedia articles needs to be represented numerically. A straightforward approach to represent text is the bag of words (BOW) model [17, 32]. In this format, every document is an observation, every word is a variable and every cell represents the presence of a word in this document. The importance of a word is traditionally expressed in one of three ways: the Term Frequency (TF), the Inverse Document Frequency (IDF) or the TF-IDF ($TF \cdot IDF$). The TF-IDF $w_{i,j}$ for term i in document j is defined as

$$w_{i,j} = tf_{i,j} \cdot log\frac{N}{df_i} \tag{1}$$

where $tf_{i,j}$ is the number of times word i appears in document j, N is the number of documents in the corpus and df_i is the number of documents in which i appears. Combining the term frequencies with the inverse document frequencies helps with the issue of term-specificity [33]. Words which appear often in many documents are overrepresented and a high term frequency of a common word is not specific to certain documents. The IDF assigns a higher weight to terms which appear only in a few documents and are hence more likely to precisely represent certain groups of documents. The TF-IDF combines both frequency of a term in a document and its specificity, creating especially high scores for rare words which appear often in one document.

After the TF-IDF feature representations are created, they are ready to be used in ML models. The traditional approach to model spatial patterns for real estate appraisal is to use regression models [21, 22, 28, 29]. Regression methods offer a simple, transparent and fast approach to most problems, but struggle when there are multicollinearity, independent variable interactions, heteroscedasticity, non-linearity or outliers present in the input data [1, 34], which is very usual for real estate data [1, 5, 13]. A modern approach, that partially solves the challenges arising from the aforementioned problems, is using ensemble learning based on decision trees as base learners. One of the most popular algorithms in ensemble learning is Gradient Boosting [35]. Each consecutively created tree is improved by being more and more negatively correlated to a predefined loss function. The goal is to iteratively draw conclusions about strong features and parameters and how to approach them to create one final and strong learner. The implementation of Gradient Boosting used in this paper is CatBoost [36]. CatBoost excels through the concept of ordered boosting, a technique which creates more randomness among the individual trees, mitigates target leakage and uses target statistics to encode categorical data.

McCluskey et al. [30] compared Gradient Boosting to linear regression on a dataset with structural variables and neighborhood membership and found Gradient Boosting to be superior in terms of predictive power. Shahhosseini et al. [22] compared the Least Absolute Shrinkage And Selection Operator (LASSO) regression, Random Forests, Artifical Neuronal Networks, Gradient Boosting and Support Vector Machines on two popular real estate datasets, namely the Boston and Ames Housing dataset. On the Boston housing data, with mostly environmental and socioeconomic variables concerning aspects surrounding the house, Random Forests and Gradient Boosting were tied for first place (similar performance for the Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE) and R^2). On the Ames housing dataset, which is mostly made up of structural variables along with one spatial membership variable, LASSO regression performed best, followed by Random Forest. Gradient Boosting came in fifth place.

3. Data Analysis

This section will give an overview of the data analysis by introducing the study area, presenting the selected dataset along with summary statistics, describing the preprocessing, introducing a distance weighted version of the TF-IDF score and explaining the tested models.

3.1. Dataset

To assess the effectiveness of Wikipedia articles for real estate appraisal, three types of data are needed: the real estate data, the geospatial data and the text data from Wikipedia. To create a baseline model from which text- and GIS-based models can be built upon, structural data for properties, along with their coordinates and sale price is needed. The dataset should contain a sufficient number of observations to ensure accuracy for ML and be from a region where GIS data is also available. The GIS data needs to contain POIs along with their category and their coordinates. We locate our experiment in Allegheny County because the available data from this region meets every aforementioned requirements. The data used is supplied by the Western Pennsylvania Regional Data Center (WPRDC) [37, 38] and contains information about houses and POIs from Allegheny County. Geolocated Wikipedia articles are available in almost every region of the USA (see Figure 2), so the corpus of geolocated Wikipedia articles can be filtered for those in Allegheny County. It should be noted that Allegheny County was chosen because of the publicly available GIS and real estate database, while the quality and quantity of Wikipedia articles was not considered in the selection process of the study area, creating a selection bias favoring the GIS models.

Allegheny County is located in the Northeast of the USA and part of Pennsylvania. It is included in the Greater Pittsburgh Region and encompasses Pittsburgh itself. It is divided by three rivers: the Allegheny River and the Monongahela River meet to form the Ohio River in Downtown Pittsburgh. As part of the Rust Belt, Pittsburgh was deemed the "Steel Capital of the World" in the industrialization era. While there is still heavy industry in Pittsburgh, it is now mainly known for its universities, museums, industrial centers and health centers.

The structural data was filtered to only include valid sales of single family homes with private owners. To control temporal effects in the local housing market, only sales from 2015 to 2020 were selected and yearly fixed effects were modeled with dummy variables. The range of sale prices was limited to $10,000 to $1,000,000, since single family homes under $10,000 seem unreasonably cheap and homes above $1,000,000 are often special cases and should be appraised on a one to one basis by a human expert. Categorical variables like roof type or building style were dummy coded, with the exception of ordinal variables, like condition or grade, which were mapped onto a cardinal scale (e.g. 1-7) since intervals between levels can, in this instance, assumed to be similar. In the end, 9,556 unique houses remained.

Table 1 offers an overview of summary statistics for the structural housing data. The mean sale price is $208,979 with a very large standard deviation of $136,597, indicating how diverse single family houses

Table 1. Summary statistics for selected numeric features representing privately owned, single family houses in Allegheny County.

Variable	Mean	Std dev	Min	Max
LOTAREA	14,953	26,747	500	897,336
SALEPRICE	208,979	136,597	11,300	996,250
STORIES	1.6	0.5	1.0	3.0
TOTALROOMS	6.6	1.4	2.0	16.0
BEDROOMS	3.1	0.8	1.0	9.0
FULLBATHS	1.5	0.7	0.0	6.0
HALFBATHS	0.6	0.6	0.0	3.0
FIREPLACES	0.5	0.6	0.0	4.0
FINISHEDLIVINGAREA	1,700	719	399	8,068

are in this dataset. The typical house has 3 bedrooms with 7 total rooms over 2 stories on 1,700 square feet of finished living area. Table 2 shows summary statistics for the 13 council districts, which will later be used for SOOS validation.

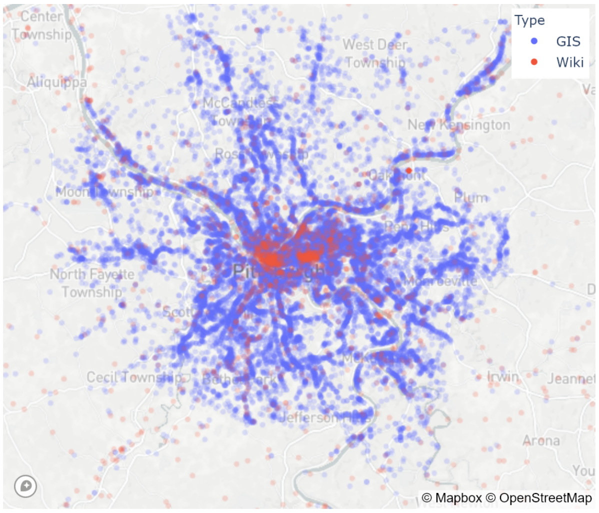

Figure 1. Visualizing geographical distribution of 23,855 POIs from the GIS database (blue) and 2,407 geolocated Wikipedia articles (red) in Allegheny County.

The GIS database was filtered to only include POIs in and around Allegheny County (see Figure 1). Duplicates were removed by selecting one of two points from the same category which were less than eight meters apart. While this may have eliminated non-duplicates (like two restaurants next to each other), eight meters will not make a significant difference for the accuracy of our features and the removal will actually improve runtime by a large margin. Categories with very few POIs, like Veteran affairs offices (5), homeless shelters (3) and WIC offices (4) were removed to limit the amount of resulting features (restricting variance) and because such a low POI count could indicate missing data. After preprocessing, 23,855 POIs remained. The most common categories are bus stop (6732), community nonprofit organization (4817), restaurant (3076), faith-based facility (953) and apartment building (908).

To create a feature set with spatial text data, an XML dump of all English Wikipedia articles was downloaded. The articles were filtered to only include geolocated articles from Allegheny County. 2,407 articles were found (see Figure 1). An NLP pipeline including tokenizing, making all words lowercase, removing stop words (including those related to Wikipedia formatting), creating 2-grams and keeping only words which appear in at least 2.2% of articles was applied. Keeping only common words helps to mitigate overfit by limiting variance and to focus on words which likely have global spatial meaning. In the end, 2,900 unique terms remained, the most common ones being ”school” (17,739), ”pennsylvania” (17,286) and ”pittsburgh” (13,005).

3.2. Weighted TF-IDF

After preprocessing the data, a suitable representation for the text needs to be found. We propose a novel spatially weighted TF-IDF approach to construct feature vectors with spatial text data for ML algorithms. Our approach incorporates the geodesic distance to each document, in this case Wikipedia articles, as well as the amount of articles which are in range of each house, for the scoring procedure.

The first modification creates spatial weights for each Wikipedia article in the TF-IDF scoring procedure. The process begins by creating traditional TF vectors for each Wikipedia article. Let K_j be the set of all TF vectors for Wikipedia articles near property j. Articles which are closer to the property should have more impact on the prediction than those further away, since they are more likely to be stronger correlated [23]. Let $d_{j,k}$ be the distance from article $k \in K_j$ to property j. The document feature vectors will be multiplied by a weighting factor which is between 0 and 1 (where, in this case, 0 is a POI d_{max} meters away and 1 is a POI 0 meters away). The standardization is global, so that every weight for every property will be equal, if it is equally far away. The d_{max} hyperparameter is important, because it determines how many articles per house are considered and works in a similar fashion to the bandwidth in GWR.

The second modification should mitigate the bias towards houses with many Wikipedia articles in range. Houses with many POIs in close proximity will have a higher TF-IDF score than some other properties for a lot of words only because there are many Wikipedia articles in range. The score should be able to characterize the document, which includes differentiating it from other documents. For a property with a very low article count, the same term frequency, or TF-IDF, of a certain word is more distinctive than for a property with an high article count. The term frequencies will hence be divided by the article count per property, creating a weighted term frequency per sub-document. In contrast to the weighted TF, which are specific to individual houses, the IDF weights are calculated on the full Allegheny County corpus, like in regular TF-IDF. The modified TF-IDF formula for word i and property j for all articles k with $d_{j,k} \leq d_{max}$ is defined as:

$$w_{i,j} = \frac{\sum_{k \in K_j} tf_{i,k} \cdot (1 - \frac{d_{j,k}}{d_{max}})}{|K_j|} \cdot log\frac{N}{n_i} \tag{2}$$

3.3. Models

To establish a **baseline**, a LASSO regression model, formalized $Linear(S+N)$, with structural features (S) along with dummy coded neighborhoods (N) is created. The neighborhoods are accounting for the location based on spatial fixed effects. This approach is traditionally often used in the literature [10, 28, 29, 30].

The **advanced baseline**, formalized $CatBoost(S+M)$, has the same feature set as the baseline, except that predictions are generated using CatBoost and dummy coded municipalities (M) were used instead of neighborhoods, since they performed better. The comparison to the baseline can show whether assumptions of linear regression hold and how results can be improved by using a more sophisticated learner.

The **GIS model**, formalized $CatBoost(S+M+D_{gis}+C_{gis})$, represents traditional approaches taken to incorporate GIS data presented in Chapter 2.2 [2, 9, 13, 29, 31] and is used as a comparison to our newly proposed approach. The POIs from the GIS database are used to calculate the geodesic distance to the nearest POI for each category (D_{gis}) and the count of POIs for each category inside a certain radius (C_{gis}). Experiments showed that a radius of 6500m for the count features was optimal.

The **WikiGIS model**, formalized $CatBoost(S+M+D_{wiki}+C_{wiki})$, tries to replicate the GIS approach with data provided only by Wikipedia. Wikipedia articles contain so called ”templates” which assign articles to groups or categories (see Table 5 for examples). These categories can be used in the same fashion as the categories from the GIS database to calculate the nearest distance to a POI for each category (D_{wiki}) and the number of POIs for each category in a certain radius (C_{wiki}). The templates were chosen by manually assessing the most common templates of geolocated articles in Allegheny County for suitability and summarizing them into distinct categories. The most important categories will be shown in Table 5.

The **Text model**, formalized $CatBoost(S+M+T)$, tries to leverage the information dense text (T) that each geolocated Wikipedia article provides by using the modified TF-IDF score described in section 3.2. Each word that appears in at least 2.2% of articles becomes a column in which the modified TF-IDF for each property is saved.

Since spatial models tend to spatially overfit by capturing too much local variance, non-useful features should be removed to mitigate overfit. Hence recursive feature elimination was performed on all models, which recursively drops the least important features for smaller and smaller sets of features and finds the optimal number of features through cross-validation. No features were removed for the baselines, while 8.0% of features for the GIS, 6.3% of features for the WikiGIS and 93.4% of terms for the Text model were dropped.

Only considering random cross-validated results can facilitate misleading conclusions, since the data shows spatial effects which should be accounted for. To test how well the proposed models perform on spatially unseen data, a SOOS validation over the 13 council districts (according to the County Council of Allegheny County) is performed. For each iteration, one district is used as a hold out test set and the model is trained on the other 12 districts. In the end, each district was used as a test set once. To gain insight into how the models perform for different types of localities, example districts along the price and the urban scale are chosen for comparison. For the price scale, two districts are selected: district 9 has the lowest average

Table 2. Average of selected variables for each council district in Allegheny County.

District	Sale Price	Lot Area	Year Built	Stories	# Houses
District 1	219,889	18,311	1959	1.6	923
District 2	347,195	28,928	1968	1.7	945
District 3	261,042	22,902	1960	1.5	1,092
District 4	191,448	16,156	1957	1.5	693
District 5	279,482	13,726	1957	1.7	1,154
District 6	159,027	10,129	1955	1.4	972
District 7	125,234	11,437	1951	1.5	809
District 8	158,731	13,535	1956	1.5	750
District 9	100,984	12,213	1951	1.3	681
District 10	190,826	7,570	1937	1.9	235
District 11	277,164	5,665	1930	1.9	419
District 12	131,674	5,992	1937	1.7	608
District 13	149,201	4,574	1926	1.9	275

price per house and therefore represents the ”affordable” category, while district 2 has the highest average price per house and thus becomes ”expensive”. As for the urban scale, average lot area per property is chosen as a proxy for urbanity. District 7 has the highest average lot area per property and is therefore an example of ”rural” whereas district 13 has the lowest average lot area per property and is consequently considered ”urban”. District 13 also has the highest average stories and buildings in this district are the oldest, further emphasizing the urbanity. Both districts along the urban scale have similar average sale price. Refer to Table 2 for more details. SOOS results can be aggregated by calculating the mean over all districts weighted by the amount of observations in each district.

3.4. Results

Table 3. Random 5-fold cross-validated results in terms of MAE and RMSE along with the standard deviation. Best results for each metric are bold.

Model	MAE	RMSE
Baseline	30,917 ± 382	45,129 ± 1,003
Adv. Baseline	30,571 ± 518	45,327 ± 1,263
GIS	28,856 ± 672	43,074 ± 1,492
WikiGIS	28,734 ± 622	42,949 ± 1,435
Text	**28,332** ± 563	**42,506** ± 1,341

The results for random 5-fold cross-validation can be seen in Table 3. The baseline and the advanced baseline achieved similar performance, with the advanced baseline slightly ahead in terms of MAE, while the baseline is ahead in the root mean squared error (RMSE). The biggest difference in performance can be observed between the advanced baseline and the more sophisticated approaches, with a 5.6% lower MAE for the GIS model in comparison to the advanced baseline. Our proposed approaches, WikiGIS and Text, are both slightly ahead of the traditional GIS approach in both observed metrics. The Text model performed best overall, outperforming the advanced baseline by 7.3% and the GIS approach by 1.8% in terms of MAE.

Table 4 shows the results for SOOS validation. The WikiGIS was again able to achieve similar performance to the traditional GIS approach. The Text model performed significantly better than all other approaches in all selected districts and the weighted mean, outperforming the advanced baseline by 13.2% and the GIS approach by 8.2% in terms of MAE, with only a tenth of spatial entities available (see Figure 1).

To show the validity of our choices for the proposed text approach, a few additional comparisons are performed, all percentage reductions are measured in terms of MAE. Removing dummy coded municipalities in random cross-validation only decreased performance by 0.4%, which indicates that the text based model learned the latent spatial patterns implicitly through text data. Omitting the TF-IDF distance weighting scheme reduced performance by 5.0% in SOOS validation when keeping the radius of 6500m, showing the effectiveness of the weighting scheme. Not using recursive feature elimination led to a decrease in performance of 4.3% for the text model in SOOS validation, indicating that too many features can lead to spatial overfit.

4. Discussion

To identify which features are relevant for quantifying location, the results described in section 3 are now further examined and compared to existing literature. For Gradient Boosting models, the impact

Table 4. MAE for SOOS validation for every model and selected council districts in Allegheny County. The last column contains a mean and a standard deviation over all districts weighted by the number of observations in each district. Best results for each district are bold.

	Districts				
Model	**affordable (9)**	**expensive (2)**	**rural (7)**	**urban (13)**	**Weighted mean**
Baseline	38,770	55,898	50,200	55,492	45,729 ± 12,408
Adv. Baseline	38,940	58,066	47,193	49,009	44,163 ± 13,390
GIS	26,563	58,273	41,932	41,932	41,785 ± 14,892
WikiGIS	35,107	**55,884**	42,293	39,258	41,686 ± 13,752
Text	**19,840**	56,242	**29,158**	**35,827**	**38,353** ± 15,622

of each feature on the prediction can be measured by feature importance. The feature importances shown in Table 5 are averages of importances gathered from location-based features of all 13 SOOS models and weighted by the number of observations in each district.

The most important location-related features for the GIS model included recreational features as well as private businesses and transportation, which is consistent with prior approaches [2, 9, 13, 29], but also introduced a new meta-category: Social Economy (affordable housing, community nonprofit organizations). The most important categories for the WikiGIS approach include accessibility (airport, transportation), education (university) and recreation (golf, park, music venue, tourist attraction), which again align with the literature [2, 9, 13, 29, 31]. It has to be noted, that the discrepancy in the most important categories for GIS and WikiGIS probably stems from the different types of categories contained in the data.

Text features are a bit more specific and show which single words had the most impact on the predictions. The most important features include words related to accessibility ("port"[1]), education ("released"[2]), recreation ("chapel") and private businesses ("commercial buildings") which are similar to the meta-categories found in WikiGIS.

Looking at the importance of single features does not give an idea of the importance of location overall. We therefore calculated the sum of feature importances for each feature group. For the advanced baseline, structural features make up 79.8% of the importance, while 20.2% are made up by dummy coded municipalities. The importance of the GIS model is divided into 64.1% structural, 5.3% municipalities and 30.6% distance & count based features. The importance of the text model is divided into 59.6% structural, 0.8% municipalities and 39.6% text based features. Unsurprisingly, structural features make up most of the sale price, though a shift can be observed in the importance of location based features. The better the feature representation for location, the less impact dummy coded municipalities have.

Sheehan et al. [18] showed that geolocated Wikipedia articles can be used to more accurately predict economic development in remote regions in Africa. This paper has shown that text data from Wikipedia articles can also be beneficial for the domain of real estate appraisal, even in data-rich and densely populated regions. In addition a different approach to create vectors from Wikipedia text data was established, demonstrating how Wikipedia articles can be more universally used, even without the usage of more complicated black box methods like Deep Learning.

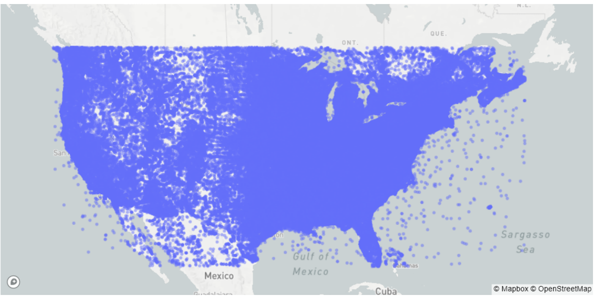

Figure 2. Wikipedia article coverage over the USA. A point is drawn for every geolocated Wikipedia article inside a rectangular box (defined by coordinates). 319,666 geolocated articles were found inside the box. Inspired by Sheehan et al. [18].

5. Conclusion

The primary objective of our study was to assess whether geolocated Wikipedia articles can be useful for real estate appraisal and compare them to traditional GIS-based methods. A novel approach was proposed that leverages text data from Wikipedia to more accurately quantify the value of location. The text

[1] The "Port Authority of Allegheny County" is the region's transit system.

[2] Articles about school districts follow a specific pattern which very frequently contains the word "released" in combination with studies and statistics.

Table 5. The ten most important spatial features for the advanced models.

	GIS	WikiGIS	Text
1.	nursing homes count	airport dist	”farms”
2.	banks count	golf dist	”port”
3.	polling places count	music venue dist	”steel”
4.	affordable housing dist	park dist	”mayor”
5.	bars count	unincorporated dist	”selected”
6.	apartment buildings count	tourist attraction dist	”chapel”
7.	bike share stations dist	church dist	”squirrel”
8.	parks and facilities dist	transportation dist	”squirrel hill”
9.	community nonprofit orgs count	park count	”released”
10.	affordable housing count	university dist	”commercial buildings”

data was processed using a modified TF-IDF formula, creating a spatially weighted vector representation. The newly devised text features outperformed traditional GIS-based features by 8.2% in terms of MAE in SOOS validation with only a tenth of spatial entities available (see Figure 1).

While results were satisfactory for Allegheny County, further research is needed to validate the results for other regions. Additionally the proposed approach should be tested against other GIS databases, since results may vary depending on the types and quality of categories. Since Wikipedia articles are crowdsourced and anyone can edit them, quality, accuracy and recency can be problematic. Real estate data can show temporal effects, due to different sale dates, which Wikipedia articles can not reflect, since they are edited for the present. Therefore temporal effects have to be accounted for. It should also be noted, that the basic TF-IDF representation does not consider word order or meaning, only the importance of each term.

Future work could either focus on further improving predictive performance with more sophisticated algorithms or increasing transparency and interpretability. To improve performance, Deep Learning techniques like attention based text transformer models [39] or word embeddings like Doc2Vec [18] could be used. To improve transparency and interpretability, incorporating techniques like LDA [40] could be helpful to study latent topics in the text and examine their impact on sale price, since single terms are volatile and not as meaningful on their own.

Because of the promising results achieved in this paper, geotagged text data should be considered alongside already established alternative data types for real estate appraisal by real estate agents and stakeholders. Wikipedia is a natural fit as a source of geolocated text, since it offers information rich and categorized text with widespread coverage (see Figure 2) and no barrier to entry [18]. Since Wikipedia articles have proven to improve predictive performance over established baselines in poverty prediction [18] and real estate appraisal, usefulness for other domains should be assessed as well.

References

[1] V. Limsombunchai, “House price prediction: hedonic price model vs. artificial neural network,” in *New Zealand agricultural and resource economics society conference*, pp. 25–26, 2004.

[2] M. De Nadai and B. Lepri, “The economic value of neighborhoods: predicting real estate prices from the urban environment,” in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 323–330, IEEE, 2018.

[3] W. McCluskey, W. Deddis, A. Mannis, D. McBurney, and R. Borst, “Interactive application of computer assisted mass appraisal and geographic information systems,” *Journal of Property Valuation and Investment*, vol. 15, no. 5, pp. 448–465, 1997.

[4] OECD, “*Tax revenues of subsectors of general government as % of total tax revenue*.” https://stats.oecd.org/Index.aspx?DataSetCode=Rev#, 2021. Accessed on 26.02.2021.

[5] S. Bourassa, E. Cantoni, and M. Hoesli, “Predicting house prices with spatial dependence: a comparison of alternative methods,” *Journal of Real Estate Research*, vol. 32, no. 2, pp. 139–160, 2010.

[6] O. Poursaeed, T. Matera, and S. Belongie, “Vision-based real estate price estimation,” *Machine Vision and Applications*, vol. 29, pp. 667–676, 2018.

[7] A. Andonov, N. Kok, and P. Eichholtz, “A global perspective on pension fund investments in real estate,” *The Journal of Portfolio Management*, vol. 39, no. 5, pp. 32–42, 2013.

[8] L. Anselin, “Gis research infrastructure for spatial analysis of real estate markets,” *Journal of Housing Research*, vol. 9, no. 1, pp. 113–133, 1998.

[9] N. Kok, E.-L. Koponen, and C. A. Martínez-Barbosa, “Big data in real estate? from manual appraisal to automated valuation,” *The Journal of Portfolio Management*, vol. 43, no. 6, pp. 202–211, 2017.

[10] S. Gröbel and L. Thomschke, “Hedonic pricing and the spatial structure of housing data–an application to berlin,” *Journal of Property Research*, vol. 35, no. 3, pp. 185–208, 2018.

[11] R. J. Hill and M. Scholz, "Can geospatial data improve house price indexes? a hedonic imputation approach with splines," *Review of Income and Wealth*, vol. 64, no. 4, pp. 737–756, 2018.

[12] C. Brunsdon, A. S. Fotheringham, and M. E. Charlton, "Geographically weighted regression: A method for exploring spatial nonstationarity," *Geographical Analysis*, vol. 28, no. 4, pp. 281–298, 1996.

[13] M. Čeh, M. Kilibarda, A. Lisec, and B. Bajat, "Estimating the performance of random forest versus multiple regression for predicting prices of the apartments," *ISPRS International Journal of Geo-Information*, vol. 7, no. 5, pp. 168–184, 2018.

[14] S. Law, B. Paige, and C. Russell, "Take a look around," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 5, pp. 1–19, 2019.

[15] C. Naumzik and S. Feuerriegel, "One picture is worth a thousand words? the pricing power of images in e-commerce," in *Proceedings of The Web Conference 2020*, WWW '20, (New York, NY, USA), p. 3119–3125, Association for Computing Machinery, 2020.

[16] J.-P. Kucklick and O. Müller, "A comparison of multi-view learning strategies for satellite image-based real estate appraisal," in *The AAAI-21 Workshop on Knowledge Discovery from Unstructured Data in Financial Services*, 2021.

[17] S. Debortoli, O. Müller, I. Junglas, and J. vom Brocke, "Text mining for information systems researchers: An annotated topic modeling tutorial," *Communications of the Association for Information Systems*, vol. 39, pp. 110–135, 2016.

[18] E. Sheehan, C. Meng, M. Tan, B. Uzkent, N. Jean, M. Burke, D. Lobell, and S. Ermon, "Predicting economic development using geolocated wikipedia articles," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2698–2706, 2019.

[19] K. J. Lancaster, "A new approach to consumer theory," *Journal of Political Economy*, vol. 74, no. 2, pp. 132–157, 1966.

[20] S. Rosen, "Hedonic prices and implicit markets: Product differentiation in pure competition," *Journal of Political Economy*, vol. 82, no. 1, pp. 34–55, 1974.

[21] W. J. McCluskey, M. McCord, P. T. Davis, M. Haran, and D. McIlhatton, "Prediction accuracy in mass appraisal: a comparison of modern approaches," *Journal of Property Research*, vol. 30, no. 4, pp. 239–265, 2013.

[22] M. Shahhosseini, G. Hu, and H. Pham, "Optimizing ensemble weights for machine learning models: A case study for housing price prediction," in *Smart Service Systems, Operations Management, and Analytics* (H. Yang, R. G. Qiu, and W. Chen, eds.), Springer Proceedings in Business and Economics, pp. 87–97, Springer and Springer International Publishing, 2020.

[23] W. R. Tobler, "A computer movie simulating urban growth in the detroit region," *Economic geography*, vol. 46, no. sup1, pp. 234–240, 1970.

[24] J. Odland, *Spatial autocorrelation*. SAGE Publications, Incorporated, 1988.

[25] P. Legendre, "Spatial autocorrelation: Trouble or new paradigm?," *Ecology*, vol. 74, no. 6, pp. 1659–1673, 1993.

[26] L. Anselin, "Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity," *Geographical Analysis*, vol. 20, no. 1, pp. 1–17, 1988.

[27] L. Anselin and D. Arribas-Bel, "Spatial fixed effects and spatial dependence in a single cross-section," *Papers in Regional Science*, vol. 92, no. 1, pp. 3–17, 2013.

[28] C. Chun Lin and S. B. Mohan, "Effectiveness comparison of the residential property mass appraisal methodologies in the usa," *International Journal of Housing Markets and Analysis*, vol. 4, no. 3, pp. 224–243, 2011.

[29] E. A. Antipov and E. B. Pokryshevskaya, "Mass appraisal of residential apartments: An application of random forest for valuation and a cart-based approach for model diagnostics," *Expert Systems with Applications*, vol. 39, no. 2, pp. 1772–1778, 2012.

[30] W. J. McCluskey, D. Z. Daud, and N. Kamarudin, "Boosted regression trees: An application for the mass appraisal of residential property in malaysia," *Journal of Financial Management of Property and Construction*, vol. 19, no. 2, pp. 152–167, 2014.

[31] I. R. Lake, A. A. Lovett, I. J. Bateman, and I. H. Langford, "Modelling environmental influences on property prices in an urban environment," *Computers, Environment and Urban Systems*, vol. 22, no. 2, pp. 121–136, 1998.

[32] S. Raschka, *Python machine learning*. Birmingham, UK: Packt Publishing Ltd., 2015.

[33] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, 1972.

[34] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112. New York, NY: Springer, 2013.

[35] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[36] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, (Red Hook, NY, USA), p. 6639–6649, Curran Associates Inc., 2018.

[37] WPRDC, "Property assessment data with parcel centroids." https://data.wprdc.org/dataset/property-data-with-geographic-identifiers/resource/2072321e-aa7c-486d-8b14-8ae79363cb68, 2020. Last accessed on 22.12.2020.

[38] WPRDC, "All allegheny county assets for the asset map." https://data.wprdc.org/dataset/allegheny-county-assets/resource/5c7825d2-6814-40c7-aefe-3d0f3d6f22e7, 2020. Last accessed on 22.12.2020.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), p. 6000–6010, Curran Associates Inc., 2017.

[40] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.