# Accessibility by public transport predicts residential real estate prices: a case study in Helsinki region

**Indrė Žliobaitė** INDRE.ZLIOBAITE@AALTO.FI
**Michael Mathioudakis** MICHAEL.MATHIOUDAKIS@HIIT.FI
**Tuukka Lehtiniemi** TUUKKA.LEHTINIEMI@HIIT.FI
**Pekka Parviainen** PEKKA.PARVIAINEN@AALTO.FI
**Tomi Janhunen** TOMI.JANHUNEN@AALTO.FI
Helsinki Institute for Information Technology (HIIT), Espoo, FINLAND

Aalto University, Dept. of Computer Science, Espoo, FINLAND

## Abstract

This pilot study investigates how considering accessibility could help to model prices of residential real estate more accurately. We introduce two novelties from the price modeling point of view (1) defining accessibility as travel time by public transport, in addition to geographic distance, and (2) considering dynamic points of interest from check-ins into social networks, in addition to fixed location community centers. Our case study focuses on the Helsinki region. We model price per square meter as a linear function of apartment characteristics, and characteristics of the neighborhood, including accessibility by public transport and social activities. The resulting models show good predictive performance, as compared to baselines not taking accessibility into account. We discover that apartment price relates to the geographical distance from the city center, but accessibility by public transport to local centers of interest is more informative than just the geographical distance to those centers.

## 1. Introduction

Modeling real estate prices has long been of interest to researchers and practitioners, and it is employed for various purposes related to investment, lending or taxation. Arguably all city residents, even non-specialists, intuitively understand that the price of a residential apartment positively relates to the size of the apartment, and negatively relates to the distance to the city center. Professional real estate price models include many more features of apartments and environment, such as age, construction type, floor, or population characteristics in the neighborhood.

Residential real estate prices are typically modeled using so called *hedonic models* (Case & Quigley, 1991; Sirmans et al., 2005), where the price of a house is assumed to be affected by the structural characteristics of the house itself, characteristics of the neighborhood, and environmental characteristics. While in real estate domain research mainly focuses on identifying factors that impact pricing, in machine learning and data mining research real estate price modeling mainly focuses on developing sophisticated predictive models beyond linear regression (Chopra et al., 2007; Fu et al., 2014).

A literature review on hedonic pricing models (Bartholomew & Ewing, 2011) finds the structural characteristics typically include the age and the size of the house, the number of bedrooms, and the presence of different amenities such as a garage. The effect of the location of the house on housing prices is often captured by physical proximity to a central business district (CBD) or a regional center. The literature review finds evidence of an inverse relationship between pricing and distance to CBD in studies on various cities around the world. Another access-related characteristic often used in hedonic models is the proximity of the house to a transit station, measured in air distance or walking distance. This attribute is used to capture the effect transit has on relative accessibility of a CBD or a regional center. Here the results are more mixed, with the majority of studies suggesting pricing premiums for housing located near to a transit station, and a higher premium for transit stations that provide a higher degree of relative proximity to a CBD.

*Proceedings of the $2^{nd}$ International Workshop on Mining Urban Data, Lille, France, 2015.* 

The era of big data provides access to new data sources, such as public transport, traffic and social mobility data, that potentially relate to real estate prices (at least intuitively we know that people consider mobility, and social factors when buying an apartment). Integrating such data could help to model residential real estate prices more precisely, and, as a result, better understand urban mobility patterns and activities. Such models can contribute to managing, coordinating and long term planning of mobility, and overall development of modern smart cities.

Our pilot study investigates to what extent accessibility of a neighborhood relates to residential real estate prices. This case study focuses on the Helsinki region. We model price per square meter as a linear function of apartment characteristics, and characteristics of the neighborhood, including accessibility by public transport and social activities. Our main hypothesis is that prices are more related to travel times than travel distances, and local centers of activities than the city center. The resulting models show good predictive performance, as compared to baselines not taking accessibility into account. We discover that an apartment price relates to the geographical distance from the city center, but accessibility by public transport to local centers of interest is more informative than just the geographical distance to those centers.

Our study introduces two conceptual novelties in modeling prices of residential real estate: (1) to measure accessibility, we consider travel times in addition to distances, and (2) we consider dynamic local points of interest, defined by 4square[1] check-ins (people posting their location and activity on a social network), in addition to community centers at fixed locations.

The remainder of the paper is organized as follows. Section 2 describes data acquisition and feature engineering. Section 3 presents the results of the experimental case study, and Section 4 concludes the study.

## 2. Data acquisition and feature engineering

Our dataset consists of three parts: real estate data describing characteristics of the apartments, location data describing points of interest and community centers, and accessibility data describing point-to-point distances and travel times. We make our dataset publicly available[2] for research.

[1] http://foursquare.com

[2] http://www.zliobaite.com/datahel.zip

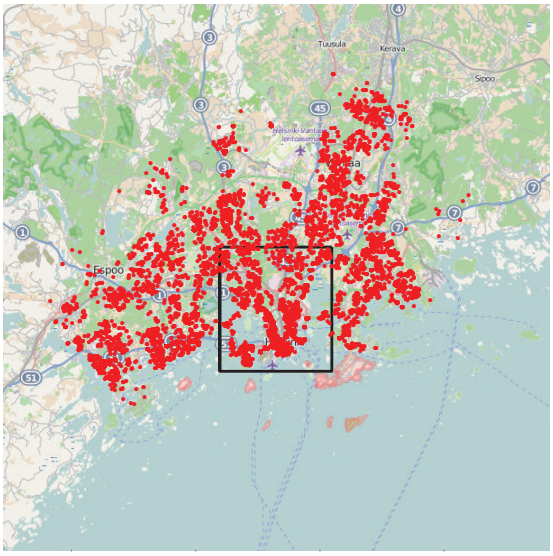


*Figure 1.* Location of apartments in the dataset. The black rectangle indicates the area from which point of interest data is collected.

### 2.1. Real estate data

The sales price data comes from a Finnish web portal Oikotie[3], which is the most popular marketplace for residential real estates sales and rental. Our dataset consists of apartments in the capital region (Helsinki, Espoo, Vantaa and Kauniainen municipalities) advertised for sales on October 24, 2014. The pricing data is based on sales ads, as sales transaction prices are not available for the public. We exclude apartments that do not provide a street address (hence no coordinates), and for which size is not available. Moreover, we filter out very large apartments (size more than 300 m2), very old apartments (built earlier than 1850), far away apartments (distance to metro more than 20 km), extremely cheap (price pr square meter less than 1200 eur) and extremely expensive apartments (price per square meter more than 12000 eur), because we aim at focusing on modeling prices of mainstream apartments and avoiding extreme outliers. After filtering our dataset includes 8337 apartments. Figure 1 plots all the apartment locations.

### 2.2. Location data

We consider two types of location data: fixed location, and dynamic points of interest. Fixed location data includes the city center, for which the Stockmann department store is used as a proxy (coordinates found by hand via Google maps), and local community centers, approximated by H&M shop (a chain of clothing shops) locations in Helsinki region (also found by hand from Google

[3] http://www.oikotie.fi/

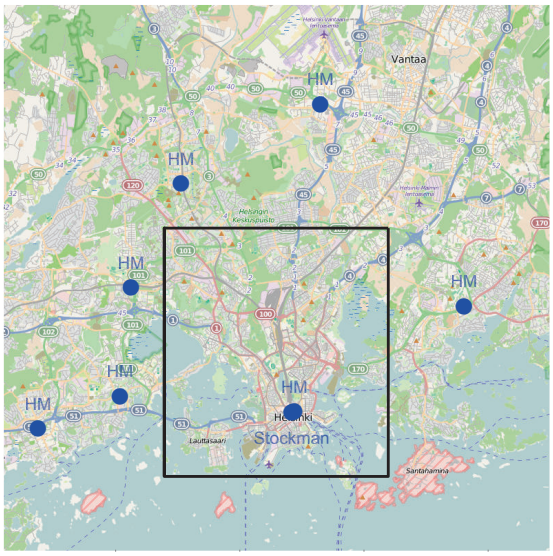


*Figure 2.* Fixed locations: community centers (H&M) and city center (Stockmann). The black rectangular indicates the area from which point of interest data is collected.

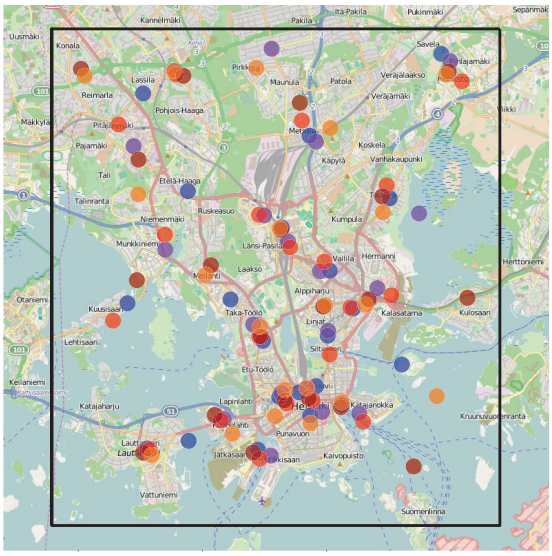


*Figure 3.* Dynamic points of interest from 4square: blue 2:00-6:00, violet 6:00-10:00, red 10:00-14:00, brown 14:00-18:00, orange 18:00-22:00.

maps). Stockmann is a well-known location in the centre of Helsinki. H&M shops are typically present in larger shopping malls. Shopping malls are local centers of attraction. We hope that H&M serves as a proxy for local centers in the neighborhoods. Figure 2 plots the community centers and the city center location.

Dynamic points of interest are obtained from an existing dataset of 4square check-ins (Le Falher et al., 2015). Each check-in in the dataset corresponds to one user's visit to one venue (restaurant, cafeteria, store, etc) with known geographic location, at a particular time. The data cover user activity between March and July 2014 in the inner Helsinki city. To extract points of interest, we perform $k$-means clustering on the geographic locations of check-ins, using $k = 20$. Each of the $k$ centroids identified defines one point of interest. Note that we extract points of interest both on top of all check-ins contained in the dataset, regardless of the the time of the day they occur, as well as separately for check-ins that occur at separate time intervals in the day (five 4-hour intervals from 2am to 10pm ). Figure 3 plots the points of interest for each time interval.

### 2.3. Accessibility features

Accessibility data connects apartments with point of interest. We consider two types of accessibility features: air distance from an apartment to the location of a point of interest, and travel time by public transport from an apartment to the point of interest (including walking time).

Air distance is measured in kilometers from coordinate of the apartment to coordinate of the point of interest, as

$$\begin{aligned} D &= R_e \cdot \arccos(s_1 + s_2), \text{ where} \\ s_1 &= \cos(lat_1) * \cos(lat_2) * \cos(lon_2 - lon_1), \\ s_2 &= \sin(lat_1) * \sin(lat_2), \end{aligned}$$

where $R_e$ is the radius of Earth (set to $R_e = 6371$km), $(lat_1, lon_1)$ are the coordinates of the apartment, and $(lat_2, lon_2)$ are the coordinates of the point of interest.

Travel time by public transport between two coordinates is measured using a freely available tool Reititin[4], developed by BusFaster Ltd and researchers at University of Helsinki. We use the default settings.

In addition to accessibility between apartments and points of interest we also include the distance from an apartment to the nearest metro station. The address of Metro stations is listed on Helsinki Metro's website[5] and their geographic coordinates are collected via manual queries to the Google Maps API[6]. Note that in the Helsinki region metro runs only to the eastern part of the city, therefore, we do not necessarily expect a regular behavior from this feature. A regular behavior would be a higher price if there is a metro stop nearby.

[4] http://blogs.helsinki.fi/saavutettavuus/tyokaluja/metropaccess-reititin/

[5] http://www.hel.fi/hki/hkl/en/HKL+Metro

[6] https://developers.google.com/maps/

*Table 1.* Input features (predictors).

| Feature | Description | Units |
|---|---|---|
| $size$ | apartment size | m2 |
| $year$ | year built | - |
| $fyear$ | function of year | - |
| $east$ | easterness | km |
| $north$ | northerness | km |
| $dmetro$ | distance to nearest metro | km |
| $ametro$ | metro within 1 km (0,1) | - |
| $dstock$ | Stockmann distance | km |
| $tstock$ | Stockmann travel time | min |
| $dhm$ | distance to H&M | km |
| $thm$ | travel time to H&M | min |
| $d4sq$ | min distance to 4square | km |
| $t4sq$ | min travel time to 4square | min |
| $d4sq1$ | distance to center 2-6:00 | km |
| $d4sq2$ | distance to center 6-10:00 | km |
| $d4sq3$ | distance to center 10-14:00 | km |
| $d4sq4$ | distance to center 14-18:00 | km |
| $d4sq5$ | distance to center 18-22:00 | km |
| $t4sq1$ | travel time to center 2-6:00 | min |
| $t4sq2$ | travel time to center 6-10:00 | min |
| $t4sq3$ | travel time to center 10-14:00 | min |
| $t4sq4$ | travel time to center 14-18:00 | min |
| $t4sq5$ | travel time to center 18-22:00 | min |

### 2.4. Final set of predictors

Altogether we consider 23 input features (predictors) for modeling apartment prices. The features are listed in Table 1.

Feature $fyear$ is a non-linear transformation of the construction year, which is based on observation that apartments built around 1970 are the least valuable, because a major pipe renovation is due in about 50 years from initial construction. Pipe renovation is done for the whole house at once, and brings substantial expenses for the apartment owners. We define the derived feature as $fyear = (year - 1970)^2$, which puts very new and very old apartments together, while apartments that are around 50 years old are put on the opposite end.

Easternness and northerness features try to capture another peculiarity of the Helsinki region, where apartments in the east are on average considered cheaper. South apartments may be considered more expensive due to proximity to the sea.

Figure 4 plots the values of selected input features against the target variable price per square meter for visual inspection. General tendencies are consistent with common intuition. The smaller the apartment, the higher the price per m2. Apartments close to metro are more expensive. The cheapest apartments have been constructed around year 1970. The most expensive apartments are in the city center, and apartments near to the local community centers or points of interest are more expensive.

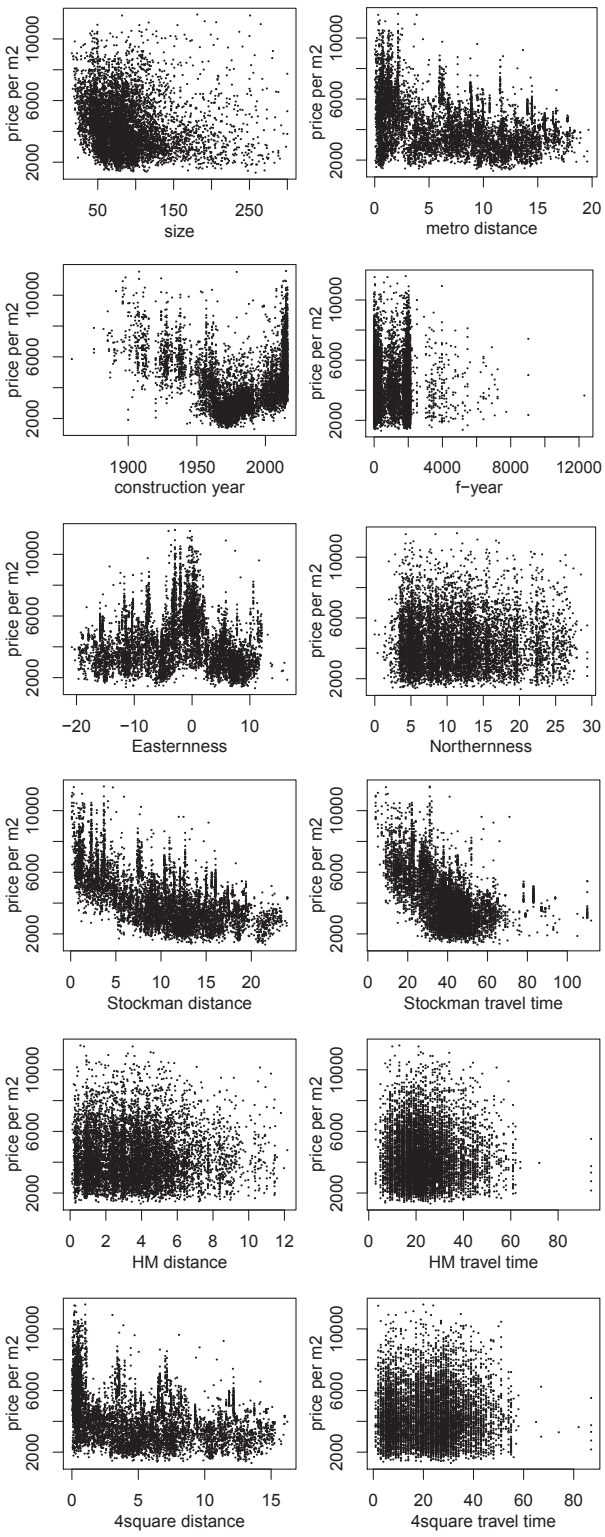


*Figure 4.* Input features against the target variable.

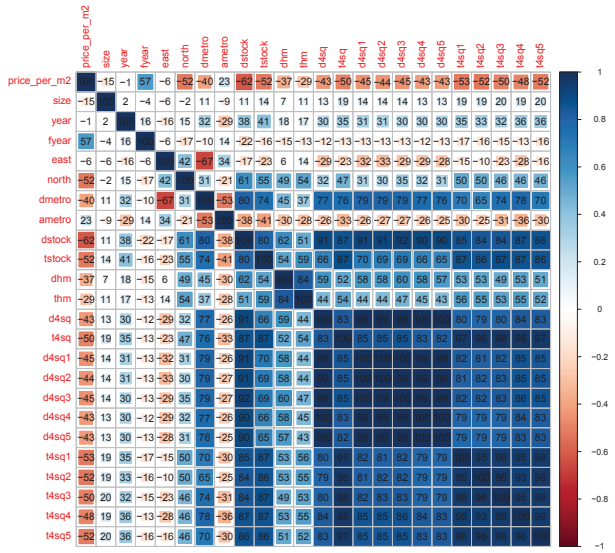


*Figure 5.* Correlations of features (a number indicates the Pearson correlation coefficient $\times 100$.

Finally, Figure 5 plots correlations across all the features. We see that the location features are strongly correlated with each other, therefore, many may be redundant. Nevertheless, some of those could potentially be expected to be more informative than others, therefore, we consider them all. We can also see than most of the location features are negatively correlated with community centers and dynamic points of interest. We already have seen similar tendencies in the scatterplots. This behavior is along with a common intuition that apartments near points of interest should be more expensive.

The correlation and scatter plots analyzed features one-by-one. In the next section we will consider predictive models that use sets of features for modeling apartment prices.

## 3. Case study

The goal of this pilot case study is to investigate whether accessibility information helps to model real estate prices, as compared to using only geographical location information. In addition, we investigate informativeness of dynamic points of interest (derived from social networks) as opposed to stationary fixed points of interest.

### 3.1. Experimental protocol

We model price per square meter. An alternative would be to model the total price. We choose the former as the target variable, because price per square meter is easier to interpret and compare across neighborhoods.

We limit our analysis to linear regression, which is easily interpretable. Note, however, that some of the features are expressed as non-linear functions of simpler features (e.g. *fyear* is a non-linear function of a building's age, as explained above). The ordinary least squares procedure (the standard implementation in R) is used for estimating the model parameters.

For assessing the performance we use two common accuracy measures: coefficient of determination (R2) and mean absolute error (MAE). Coefficient of determination is a relative accuracy measure, where 1 means the best possible performance, and 0 means the performance is equivalent to random. Mean absolute error indicates error in the units of the target variable, 0 is an ideal performance, the higher the MAE, the worse the performance.

We report R2 and MAE measured on the whole dataset used for model fitting (fit) and via 10 fold cross-validation (cv), which iteratively fits a model on 90% of the data, and tests on the remaining part. Cross-validation scores provide an indication of how models would generalize to unseen data.

### 3.2. Performance of base models

Base models do not use any accessibility information, and use only very basic location information. The first base model (Size-year) does not use location at all, and is based only on size of the apartment and its construction year (*fyear*). The second model in addition uses basic location information, encoded as raw geographical coordinates, centered in the old town of the city.

The resulting models for price per square meter are:

$$price = 3722 - 4.97 \times size + 0.91 \times fyear,$$

and

$$\begin{aligned} price &= 5643 - 5.14 \times size + 0.78 \times fyear + \\ &+ 38.9 \times east - 147.7 \times north. \end{aligned}$$

The models are consistent with common intuition: the larger the apartment, the cheaper the price per square meter; older or newer apartments with respect to 1970 construction year are more expensive; the further to the north from the sea and the city center, the cheaper. Easterness has a positive effect, which is somewhat inconsistent with a common intuition that cheaper neighborhoods are in the east. However, this can be explained by the range of data (see Figure 4). Data extends further to the west than to the east, therefore, western apartments are on average further from the center, and thus cheaper.

Table 2 reports predictive accuracies of the base models. We can make two observations. First, Size-year-location model already performs quite well with the cross-validation

*Table 2.* Base models for predicting price per m2. R2 - coefficient of determination (the higher, the better), MAE - mean absolute error (the smaller the better).

| Model | R2 fit | MAE fit | R2 test | MAE test |
|---|---|---|---|---|
| Size-year | 0.34 | 1062 | 0.33 | 1063 |
| Size-year-location | 0.56 | 836 | 0.55 | 837 |

*Table 3.* Accessibility for predicting price per m2. Size-year-location model with one additional accessibility feature at a time.

| Add feature | R2 fit | MAE fit | R2 test | MAE test |
|---|---|---|---|---|
| Metro distance | 0.58 | 811 | 0.58 | 813 |
| Metro access | 0.56 | 835 | 0.55 | 836 |
| H&M distance | 0.56 | 837 | 0.56 | 838 |
| H&M travel time | 0.56 | 831 | 0.56 | 832 |
| Stockmann distance | 0.61 | 781 | 0.61 | 782 |
| Stockmann t. t. | 0.58 | 811 | 0.58 | 812 |
| 4square dist. all | 0.58 | 813 | 0.58 | 814 |
| 4square t. t. all | 0.58 | 804 | 0.58 | 805 |
| 4square dist. peak | 0.59 | 807 | 0.59 | 809 |
| 4square t. t. peak | 0.59 | 799 | 0.59 | 800 |

R2 result 0.55. Second, the fit and the cross-validation performance differs only a little, which suggests that there is no notable overfitting, and the model could use more informative input features.

### 3.3. Predictive power of accessibility

Next we test whether adding accessibility information helps to predict more accurately. We test accessibility to the city center (Stockmann), local centers (HM), and dynamic centers of interest (4square check-ins) overall and at morning peak times (from 6:00 to 10:00). We compare informativeness of using air distance as a feature to using the total travel time by public transport.

We use the base model Size-year-location as a starting point, add one feature at a time to it, and measure the accuracy. Table 3 reports the results.

From the resulting accuracies we can see that accessibility has some predictive power, as in all cases the predictive performance improves as compared to the base model. The results indicate that the distance to the city center (Stockmann) is more informative than the travel time by public transport. However, accessibility to the local centers (fixed centers H&M and dynamic centers 4square) by public transport is more informative than just the air distance to those centers. In other words, it seems that an apartment price relates to the overall geographical location, but *accessibility* to local centers of interest is more important than just the geographical distance to those centers. This is an interesting finding for exploring in detail in future studies.

We report selected models. Metro distance is intuitive - the closer to metro, the more expensive is the apartment:

$$\begin{aligned} price &= 5300 - 3.94 \times size + 0.77 \times fyear - \\ &- 62.2 \times east - 50.8 \times north - 158.4 \times metro. \end{aligned}$$

Adding metro distance shrinks other coefficients, which suggests that earlier this feature was indirectly captured. More importantly, adding metro distance changes the direction of the easternness coefficient from positive to negative. Now it is more intuitive keeping in mind peculiarities of Helsinki residential neighborhoods, where overall the east is considered cheaper than the west.

Stockmann distance is as well intuitive - the closer to the center, the more expensive is the apartment:

$$\begin{aligned} price &= 5698 - 3.78 \times size + 0.72 \times fyear + \\ &- 3.3 \times east - 59.8 \times north - 117.8 \times dstock. \end{aligned}$$

Shortest H&M travel time is intuitive - the shorter the travel time to the local center, the more expensive is the apartment:

$$\begin{aligned} price &= 5681 - 5.00 \times size + 0.78 \times fyear + \\ &+ 37.0 \times east - 139.6 \times north - 39.3 \times thm. \end{aligned}$$

Shortest 4square travel time is intuitive - the shorter the travel time to a center of interest, the more expensive is the apartment:

$$\begin{aligned} price &= 5659 - 3.55 \times size + 0.77 \times fyear + \\ &+ 13.5 \times east - 103.3 \times north - 31.4 \times t4sq. \end{aligned}$$

### 3.4. Final predictive model - everything together

Finally, we collect a set of promising features into one final model, and test its performance. The final model includes the base model (Size-year-location), metro distance, Stockmann distance (a proxy for distance to the city center), and travel times to H&M and 4square (peak) local centers.

While the final feature selection is done after seeing the intermediate performance results, the fit accuracies have been very similar to those of cross-validation, therefore the risk of overfitting is not high.

The final model fitted on all the data is

$$\begin{aligned} price &= 5729 - 4.06 \times size + 0.71 \times fyear + \\ &+ 31.6 \times east - 94.5 \times north + \\ &+ 73.0 \times metro - 139.0 \times dstock + \\ &+ 21.6 \times thm - 12.5 \times t4sq2 \end{aligned}$$

We can see some interesting relations, reflecting peculiarities of the Helsinki region. First, the longer the metro

*Table 4.* Final model for predicting price per m2.

| Model | R2 fit | MAE fit | R2 test | MAE test |
|---|---|---|---|---|
| Final model | 0.63 | 752 | 0.62 | 753 |

distance, the higher the price, while one could expect the opposite. Our interpretation is that the metro distance captures what was not very successfully captured by the easternness. We observe that the coefficient of easternness shrinks when metro comes into the equation. In Helsinki metro runs only to the eastern suburbs, and these suburbs are considered less prestigious neighborhoods than the west, and, therefore, residential prices there are lower.

Table 4 reports the performance figures. The final model shows the best performance seen so far, and reasonably good accuracy in relative and absolute terms (testing R2 = 0.62).

## 4. Conclusion

We have experimentally explored several models for real estate prices in Helsinki region, focusing our analysis on accessibility by public transport and dynamic points of interest, obtained via check-ins into social networks. We have found that even a basic account for accessibility features helps to improve the accuracy of price estimates. We have discovered that an apartment price relates to the geographical distance from the city center, but accessibility by public transport to local centers of interest is more informative than just the geographical distance to those centers.

Integrating such data could help to model residential real estate prices more precisely, and, as a result, better understand urban mobility patterns and activities. Such models can contribute to managing, coordinating and long term planning of mobility, and overall development of modern smart cities.

## Acknowledgments

The authors thank Antti Ukkonen for insightful discussions. Research leading to these results was partially supported by the Aalto University AEF research programme and the Academy of Finland grants 118653 (ALGODAN) and 251170 (Finnish Centre of Excellence in Computational Inference Research COIN). Maps are credited to ©OpenStreetMap contributors, for more information see http://www.openstreetmap.org/copyright.

## References

Bartholomew, Keith and Ewing, Reid. Hedonic price effects of pedestrian and transit-oriented development. *Journal of Planning Literature*, 26(1):18–34, 2011.

Case, Bradford and Quigley, John M. The dynamics of real estate prices. *The Review of Economics and Statistics*, 73(1):50–58, 1991.

Chopra, Sumit, Thampy, Trivikraman, Leahy, John, Caplin, Andrew, and LeCun, Yann. Discovering the hidden structure of house prices with a non-parametric latent manifold model. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pp. 173–182, 2007.

Fu, Yanjie, Xiong, Hui, Ge, Yong, Yao, Zijun, Zheng, Yu, and Zhou, Zhi-Hua. Exploiting geographic dependencies for real estate appraisal: A mutual perspective of ranking and clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pp. 1047–1056, 2014.

Le Falher, Geraud, Gionis, Aris, and Mathioudakis, Michael. Where is the Soho of Rome? : Measures and algorithms for finding similar neighborhoods in cities. In *The 9th International AAAI Conference on Web and Social Media*, ICWSM, 2015.

Sirmans, Stacy, Macpherson, David, and Zietz, Emily. The composition of hedonic pricing models. *Journal of Real Estate Literature*, 13(1):1–44, 2005.