Effects of transport systems on housing rents

Maurin Baillif

Matthieu de Lapparent

IIDE - HEIG-VD : School of Business and Engineering Vaud

May 2020

STRC | 20th Swiss Transport Research Conference
Monte Verità / Ascona, May 13 – 15, 2020

IIDE - HEIG-VD : School of Business and Engineering Vaud

Effects of transport systems on housing rents

Maurin Baillif, Matthieu de Lapparent
IIDE - HEIG-VD : School of Business and
Engineering Vaud
HES-SO - University of Applied Sciences and
Arts Western Switzerland
Avenue des sports 20 - CP 521 - 1401
Yverdon-les-Bains, Switzerland.
phone: +41-24-557-75-39
fax:
{maurin.baillif,matthieu.delapparent}@heig-vd.ch

May 2020

Abstract

Accurate estimation of property prices is a key issue for both households and companies. Any household interested in buying or renting a dwelling will focus on some sort of price-quality ratio (here embedding intrinsic and extrinsic values of the characteristics of the proposed dwelling). Any intermediate real estate company is optimizing her asset's portfolio or practicing vertical differentiation by proposing new approaches. We nonetheless observe that the latter often builds up on very standard valuation methods (capital value, residual value, intrinsic value, discounted cash flow, hedonic prices to a lesser extent) to minimize risk of error in property valuation. Focusing on the hedonic pricing approach in an empirical framework, we here focus on the role and effects of transport systems and ways they are determining dwelling values & rents at a local level. From cumulated opportunities using simple crow-fly distances within given radius to more complex indicators using general cost, we here show that there exists very local circumstances in using one or another of those indicators for market valuation. From a methodological perspective, we propose a supervised discrete mixtures of market hedonic pricing structures. We implement an Simulated Expectation-Maximization algorithm with accuracy criterion and compare it with the classical hedonic pricing model. Our baseline application focuses on rent data. These are spatially augmented using open source data. Results show that use of our approach improves precision of predicted rents. Shapes and roles of accessibility measures are also easy to spatially understand.

Keywords

Hedonic regression, Latent class, Simulated Expectation-Maximization algorithm

1 Introduction

Price prediction is an important topic in real estate. It helps to build and manage real estate investment portfolios, drive urban development policies, advise to individuals on purchasing or selling, and defend interests of various stakeholders (landlords, tenants, banks, investment funds). Today prices and rents in Switzerland are often established basing on financial models that do not take into account (or take into account in a very limited way) attributes of the housing, its environment and evolution of local and global economic conditions. The hedonic method is a well-known alternative which takes into account attributes for the price estimation.

This method starts from a simple assumption: the price of the good can be considered as the sum of the prices of its characteristics (Rosen, 1974). Therefore, the price difference between two goods is explained by differences in the number of their characteristics and the marginal price (or hedonic price) of each of these characteristics. The hedonic price is calculated through the linear regression between the prices of goods and the characteristics of these goods. Most of the research suggests to use the hedonic pricing method on a area or country assuming it is homogeneous. Consequently, a characteristic common to all goods (e.g. number of rooms) explains in identical proportions the final price of all goods. However, we have the intuition that not all real estate goods in a region, or country, are evaluated in the same way and therefore on the same characteristics or the same hedonic prices. For example, some goods will be more sensitive to certain characteristics such as road noise or accessibility to public transport, while others will be more sensitive to floorspace or number of rooms. These differences could be explained by the type of property (standard property vs. exceptional property), its location (urban vs. rural) etc.

The theoretical framework of the hedonic pricing method is flexible, there are no particular constraints on the functional form or the choice of variables to be used. In this paper we will take advantage of this flexibility by assuming that, for the same area as above, it is possible to highlight different groups (or classes) behaving differently using latent class (Lazarsfeld, 1950).

Focus on the impact of public transport accessibility and applied on a real estate dataset, the objectives of this paper are multiple. First, the standard hedonic pricing method is used to determine the model that best describes prices, according to different accessibility variables. One counting opportunities as the crow flies and the other counting opportunities using generalized cost. Then, based on the previously selected model, the Simulated Expectation-Maximization algorithm (SEM) is used to determine the membership of each dwelling to one of the latent classes and the associated coefficients. Finaly, the accuracy of each class and the overall model will be measured based on root-mean-square error criterion (RMSE) and compared with the RMSE of the standard hedonic pricing method.

The rest of the paper will be structured as follows, Section 2 presents the literature review. Section 3 describes the mathematical framework that will be used. Section 4 describes the case studies, a real estate announcement dataset in Western Switzerland. Section 5 presents and discusses the results of the case studies according to the methodology presented. And, finally, Section 6 concludes on the main outcomes of the research and the future work.

2 Literature review

Hedonic price analysis methods are often associated with Rosen (1974). His paper was the first to set the theoretical structure to hedonic pricing. Rosen defined market equilibrium conditions for the short and long term and suggested a two-step procedure for estimation of housing demand function. The first step is regression of prices on observable attributes of the good and computation of implicit marginal prices in each point. This step is also known as market valuation of housing characteristics. It is the approach that is mostly used in empirical analysis. The second step is using estimated marginal prices as endogenous variable in the second-stage estimation of housing demand functions. However, studies on the price definition of complex goods were made before. The very first work on the topic was made by Waugh (1929) about the analysis of vegetables prices as a function of characteristics related to their qualities. Later, Court (1939) proposed the first formal contributions to hedonic price theory with automotive examples.

Hedonic pricing received widespread usage, applied research dedicated to analysis of real estate prices has developed, driven both by interest to the knowledge of the sector, but also by economic problems and econometric issues faced. Studies of housing attributes' pricing often raise complex theoretical and statistical problems since these prices are not directly observable on market (Ellickson, 1981). They are needed to be derived from a different price, one of the housing. Fondamental hypothesis is that a housing good and its location are inseparable, they are sold together, and consumer transforms them to get utility.

Since there are no specifications regarding the functional form of the model to be used, there is no consensus among researchers on this subject. Admittedly, in most cases the log-linear model is used insofar as it easily allows the integration of explanatory variables and an easy interpretation of coefficients. Some authors justify their choice by the decrease in heteroskedasticity with the use of the logarithmic form Picard *et al.* (2012). Linear and log-log models are also regularly used (Cavailhès, 2005) ; however, we have only identified one case where the lin-log configuration has been used, along with other combinations of the linear and logarithmic terms (Melichar *et al.,* 2019).

Nor is there any consensus on the variables to be used and their form. The explanatory variables used can be continuous in linear, quadratic or logarithmic form but also discrete (Ruf, 2017, Cavailhès *et al.*, 2009, Boucq and Papon, 2008). Some studies use the same data several times by transforming it (Cebula, 2009). A great diversity as well as a certain evolution over time are observed. Gross *et al.* (1990) uses 12 variables when, 20 years later, Cavailhès *et al.* (2009) uses more than 30. In addition, extrinsic explanatory variables such as noise, view or distance to public transport have been included in the hedonic model (Löchl and Axhausen, 2010, Baranzini *et al.,* 2008, 2010). We will use, among other things, the results of these different studies to select and construct the variables used in the Sections 4 and 5.

In the mean time, machine learning (ML) techniques, which have been extremely popular in recent years, have obviously been tested on subjects related to the estimation of real estate prices. There are several publications using ML techniques with "black box" effects (Oladunni and Sharma, 2016, Wezel *et al.*, 2005, Mayer *et al.*, 2018), focusing on accuracy. But very few use the SEM algorithm with linear models as an approach to highlight latent classes (Howell and Peristiani, 1987) and use the obtained results to calculate hedonic prices of the attributes of a good.
The EM algorithm is one of these machine learning techniques. It is an iterative algorithm first proposed by Dempster *et al.* (1977) to define the parameters of the maximum true likelihood of a probabilistic model when the latter depends on unobserved latent variables. This algorithm is a general method for maximum-likelihood estimation. Its alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step. Since its inception, many applications have used the SEM algorithm, both in the field of environmental science (Mustafaa *et al.*, 2011), computer science (Liu *et al.*, 2006, Han and Xue, 2015), medical science (Yang *et al.*, 2012, Fraley and Raftery, 2002) or financial science (Gorshenin *et al.*, 2017).

Among all the concepts presented in this paper, that of accessibility is certainly the oldest and paradoxically the most complicated to approach. Moroz (2017) defines it as the combination of two components, specific to each study. A *transport* component, refering to the distance, travel time, travel costs, travel effort, or more generally, the perception and valuation of this time and effort on the part of a traveler. And a *land-use* component, refering to the spatial distribution of demand for activities or the "opportunities" available in an area. Many models exist and some studies have attempted to compare them (Song, 1996, Geurs and van Wee, 2004).
In the real estate context of this paper, the concept of accessibility is interpreted as the accumulation of opportunity, represented by public transport stops, according to the effort it takes to get there from the location of the property. In order to study their differences, the choice is made to keep two models with different philosophies. The Euclidean model is one of the simplest measures of accessibility to calculate (Vickerman, 1974). Accessibility indicator correspond to the accumulation of public transport opportunities in a given radius. For instance, 400 meter buffers around bus stops and 800 meters around rail stations are commonly used to identify the area from which most transit users will access the system by foot (El-Geneidy *et al.*, 2009). The gravity model (Hansen, 1959), is the most common approach. The idea is to associate it with the

concept of generalised cost so that the attractiveness of a public transport stop decreases both with the increase in distance and gradient.
The real estate sector has also been the subject of studies to examine the relationship between the supply of transportation in a region and real estate prices (Martinez and Viegas, 2009, Chen *et al.*, 2019). These approaches use the standard hedonic princing method, and therefore assume that the properties behave identically.

This paper proposes to go further by studying accessibility while combining hedonic pricing model and SEM algorithm. The purpose of this association is to highlight the latent class structure of the real estate market, study the differences inside Latent classes andevaluate the accuracy gain of such an approach compare to standard hedonic price model.

3 Methodology

This section presents the methodology that will be used in the rest of the paper. The first step will be to perform hedonic regression using the ordinary least squares method to estimate the best performing model, regarding accessibility indicators. This method, which is widely known, will not be presented in this section. The methodology focuses on the accessibility indicator definition and the evaluation of the best model using the Simulated Expectation-Maximization (SEM) algorithm, followed by a criterion measuring the accuracy of the model estimation.

3.1 Accessibility indicator definition

Euclidean distance

A simple way to define accessibility is to count everything within a certain distance as the crow flies. Therefore, the first accessibiliy indicator A_i^1 can be express as :

$$A_i^1 = \sum N_j \; I(d_{i,j} < \overline{d}) \tag{1}$$

With N_j the number of opportunities at location j , $d_{i,j}$ the distance, in meters, between the dwelling and the bus stop, the subway or the train station and $\overline{d}$ a specific radius (200 m / 400 m / 600 m and 800 m).

The limit for this accessibility indicators, based solely on distance, comes into play when the slope rises. Some cities in French-speaking Switzerland such as Lausanne or Neuchatel are known to have a steep gradient.

Gravity Model

The Gravity model aims to improve the Euclidean model by taking into account topographic effects. In this model, the effort to go from i to j is measured by a *Generalized Cost* and structured so that accessibility decreases as distance increases. The choice of the functional form being free, we choose the negative exponential, which is the most common functional form and the most reflective of actual behavior. Therefore, the second accessibiliy indicator A_i^2 can be express as :

$$A_i^2 = \sum N_j \exp^{-\lambda CG_{i,j}} \tag{2}$$

With, N_j the number of opportunities at location j, λ a sensitivity parameter fixed to 1 and $CG_{i,j}$

the generalized cost between i and j.

This generalized cost is defined using the notion of "effort kilometer" (*kme*), used for mountain trails. The *kme* transforms the notion of elevation gain into kilometres. This makes it possible to approximate a distance with a difference in altitude to what it would be on a flat surface. Generally, the kme is defined as follows :

$$kme_{i,j} = \frac{d_{i,j}}{1'000} + \frac{\sum D^{+}_{i,j}}{100} + \frac{\sum D^{-}_{i,j}}{300} \tag{3}$$

With $D^{+}_{i,j}$ the positive altitude difference and $D^{-}_{i,j}$ the negative altitude difference. A differentiation is made between a positive and a negative gradient in the sense that, at the same distance, the uphill climb requires greater effort than the downhill descent. In this context, the assumption is made that an individual assesses accessibility solely on the basis of uphill distance. Therefore, the second accessibility indicator can finally be defined as :

$$A^2_i = \sum N_j \exp^{-\lambda(\frac{d_{i,j}}{1'000} + \frac{\Delta h_{i,j}}{100})} \tag{4}$$

3.2 Simulated Expectation-Maximization (SEM) Algorithm

The SEM algorithm is an approach used to perform maximum likelihood when it depends on unobservable latent variables. It is composed of 2 steps: an expectation evaluation step (E), where the expectation of likelihood is calculated taking into account the last observed variables and a maximization step (M), where the maximum likelihood of the parameters is estimated by maximizing the likelihood found in step E. Under these conditions, the algorithm repeats its two steps in an iterative way in order to converge towards maximum likelihood.

For a defined population, each individual i belongs to one and only one of the k classes. The price of a real estate good y_i can therefore be defined by the usual hedonic price equation by making it depend on the membership of its class k:

$$y_i = x_i^{'}\beta_k + \epsilon_{i,k} \tag{5}$$

With $x_i^{'}$ the characteristics of a real estate property, β_k the regression parameters and $\epsilon_{i,k}$ the error term, $\epsilon_{i,k} \sim \mathcal{N}(0, \sigma^2_k)$. For this step, variables $x_i^{'}$ are the same for all classes.

We assume that prices y_i of each class follow a normal distribution with the following probability density :

$$\Pr(y_i|X_i,\beta_k,\sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{1}{2}\frac{(y_i - x_i^{'}\beta_k)^2}{\sigma_k^2}} \tag{6}$$

The class membership can not be directly observed, then, distribution probability of each class is express as :

$$\Pr(y_i|X_i,\beta_1,\delta_1,\sigma_1^2,\ldots,\beta_k,\delta_k,\sigma_k^2) = \sum_{k=1}^{K} \delta_k \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{1}{2}\frac{(y_i - x_i^{'}\beta_k)^2}{\sigma_k^2}} \tag{7}$$

With $\delta_1, \ldots \delta_k$, $\delta_k \in [0,1]$ and $\sum \delta_k = 1$.

The full information maximum likelihood function of the problem is defined as follows :

$$ll = \sum_{i=1}^{n} ln\left(\sum_{k=1}^{K} \delta_k \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{1}{2}\frac{(y_i - x_i^{'}\beta_k)^2}{\sigma_k^2}} \right) \tag{8}$$

In order to avoid dealing with complex non linearities we prefer to build on the SEM likelihood function. It is defined as :

$$ll_{SEM} = \sum_{i=1}^{N} \sum_{k=1}^{K} d_{i,k}^{t} \Big[ln(P(y_i|X_i,\beta_k^t,\sigma_k^{t^2})) + ln(\delta_k^t) \Big] \tag{9}$$

With, $d_{i,k}$ a random vector multinomially distributed and t the iteration.

The SEM algorithm is used in order to estimate the values δ_k, β_k and σ_k. As explained, this iterative process is composed of an Expectation step (E-step), a Multinomial distribution step and Maximization step (M-step).

E-step :

The objective is to estimate unknown parameters $\gamma_{i,k}$ of t iteration. $\gamma_{i,k}$ represent the probability that an individual has to belong to a class k. Values are estamted using parameters δ_k, β_k and σ_k

of the previous iteration, $t-1$.

$$\mathbb{E}(d_{i,k}|y_i) = \gamma_{i,k}^t = \frac{\delta_k^{(t-1)} \frac{1}{\sqrt{2\pi\sigma_k^{(t-1)2}}} e^{-\frac{1}{2}\frac{\left(y_i - x_i'\beta_k^{(t-1)}\right)^2}{\sigma_k^{(t-1)2}}}}{\sum_{k=1}^K \delta_k^{(t-1)} \frac{1}{\sqrt{2\pi\sigma_k^{(t-1)2}}} e^{-\frac{1}{2}\frac{\left(y_i - x_i'\beta_k^{(t-1)}\right)^2}{\sigma_k^{(t-1)2}}}} \tag{10}$$

With $\sum_{k=1}^K \gamma_{i,k} = 1$.

Multinomial-step :

During iteration t, a random vector $d_{i,k}$ multinomially distributed is defined using values $\gamma_{i,k}$ of step E. With $d_{i,k} \sim \mathcal{M}(1, \gamma_{i,k})$.

M-step :

The log-likelihood is maximized, using values of the previous step for Eq. (9). The maximization step then solves, for each class, the following conditions. The beta coefficients of the regressions :

$$\hat{\beta}_k^t = \left[\sum_{i=1}^N d_{i,k}^t X_i X_i'\right]^{-1} \left[\sum_{i=1}^N d_{i,k}^t y_i\right] \tag{11}$$

The standard deviations of regressions :

$$\hat{\sigma}_k^{2,t} = \frac{\sum_{i=1}^N d_{i,k}^t (y_i - X_i'\beta_k^t)^2}{\sum_{i=1}^N d_{i,k}^t} \tag{12}$$

The weighting of each class :

$$\hat{\delta}_k = \frac{\sum_{i=1}^N d_{i,k}^t}{n} \tag{13}$$

3.3 Model accuracy criterion

Root-mean-square error criterion (RMSE) measures the accuracy of a model using both predicted and true values. The classic formula is adapted in order to be used with a latent class model.

$$RMSE_{SEM} = \sqrt{\frac{\sum_{i=1}^{N} \sum_{k=1}^{K} d_{i,k}(y_{pred,i,k} - y_i)^2}{N}} \quad (14)$$

With $d_{i,k}$ the belonging of an observation i to a class k, N the overall sample size, y_i the real value of an observation, $y_{pred,i,k}$ the predicted value of an observation for a class k.

This criterion will make it possible to compare RSME for a standard hedonic regression and hedonic regression using the SEM algorithm.

4 Case study

This section presents the real estate announcements dataset that is used to apply the methodology presented in Section 3.

4.1 Variable description

Present approach is built up on combining several sources of open data. Data on houses and their key attributes are extracted from announcements on Internet. The prototype of the model was created and tested using a sample gathered on 1st and 2nd November 2017. Raw data contain 6634 observations for which exact address, geographical coordinates, floor space, number of rooms, canton and rent amount were collected. Objects with surface above 500 square meters, as well as 0.5% of the least and the most expensive dwellings were removed from the sample.

Dependent variable used in analysis is monthly rent per square meter. It is important to stress that rent amount is as it is stated in the announcement, and not necessarily rent that tenant will pay. It means that the endogenous variable is subject to measurement error. For example, in Switzerland tenant can contest the rent after concluding the contract. However, we assume that in general most of rents published are final rents. Throughout this document, and by misuse, we will regularly use the notion of "price" to refer to the "rent".

The Fig. 1 shows the location and the price over floorspace of all real estate properties in the different cantons of Western Switzerland. We can see that the data are well distributed over the entire territory with a greater concentration around the Lake Geneva region and the price over floorspace seems higher in cities (eg. Lausanne and Geneva area). These comments are consistent with the large population pool of the Lake Geneva region and the fact that rents in the cities are higher than in the countryside.

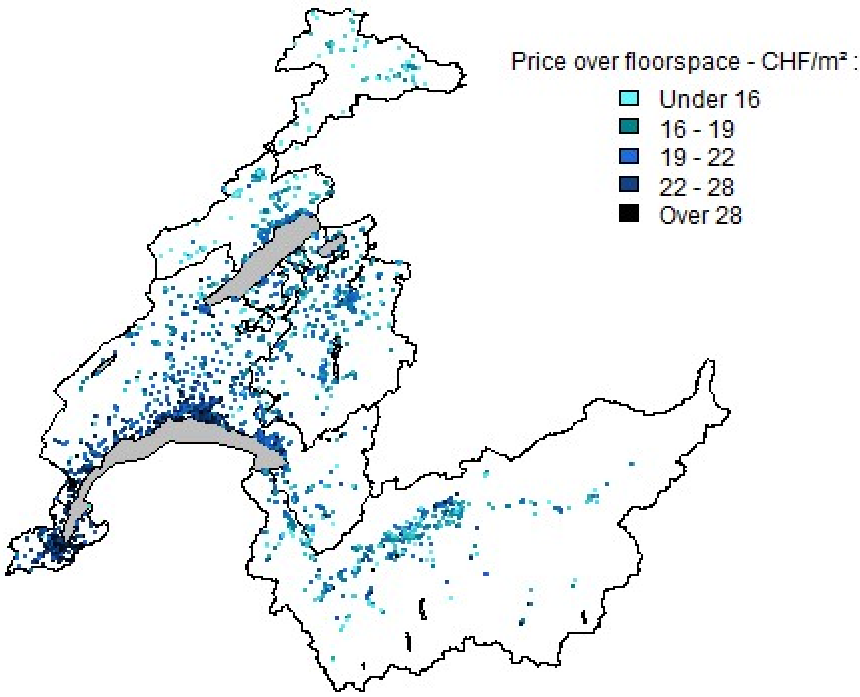

Figure 1: Location of real estate property

In order to enrich the data collected on the Internet, new variables have been created from different open data sources.

There is one variable on housing itself that is not present in announcements: age of building. Its was extracted from StatBL survey made by FSO. These are georeferenced data with grid 100x100 meters. In other terms, data available show average age of buildings that occupy a given cell. Being unable to define precisely the age of the building, we still dispose a measure of age for the closest neighbourhood.

Employment, population, household size and age of buildings data were taken from Federal Statistical Office (FSO).More precisely, employment from STATENT-2015 surveys, population and household size from STATPOP-2016 survey and age of buldings from StatBL-2016. All the information extracted from these data are aggregated at hectare level.

Source for information on noise is SonBase data from Federal Office for Environment (FOEN). Dataset on route noise is from 2010, and the one on railways is from 2011. These data deliver exposure to road and rail noise sources in a georeferenced 10x10m grid.

Finally, accessibility indicators are also created using a 2016 "Public transport stops" database, taken from Federal Statistical Office (FSO). The definition of accessibility indicators is presented in Section 3.1.

The years of data collection of these different open data are not the same. We assume that this will have only a very slight impact on the results since the orders of magnitude are representative and there have been no major changes (demographic or economic) in recent years.

There is still some important information on housing features that has not been collected like the presence of parking slot, the view or vacancy rate. We have chosen to limit the number of variables used while diversifying the information about the properties studied: intrinsic data (floorspace, number of rooms) vs. characteristics of environment (noise, accessibility). The variables used to explain the price definition are presented in Table 1.

Variable	**Description**
Floorspace	Log of the surface, in m^2
Number of rooms	Number of rooms in the real estate property
Employment number	Log of the employement number of workplaces (full-time equivalent) in a given radius (10 km)
Population number	Log of the population number of workplaces (full-time equivalent) in a given radius (10 km)
Household size	Log of the average household size within a hectar
Age of buildings	Log of the average age of buildings within a hectar
Road noise	noise caused by the road at night, in dB
Train noise	noise caused by the train at night, in dB
Canton	Canton where the object is located:VD, VS, GE, JU,FR (reference : FR)
Accessibility indicator	We defined two indicators, one based on the distance from the dwelling to public transport stops, the other integrating both the distance and the difference in altitude from the dwelling to public transport stops

Table 1: Variable description

4.2 Descriptive statistics

The descriptive statistics of the dwellings are presented in Table 2. Basically, for the Western Switzerland, the average household consists of 2.5 people, living in a dwelling of $87m^2$, with 3,5 rooms and located in a district where the age of the buildings is 54 years. The noise from the road is 40 db and 14db from the train.

Within a 10 km radius, the average population is 130'000 and employment opportunities are 65'000. Intuitively, we imagine that this average property is located on the periphery of a large city.

Finally, Vaud is the most represented canton with 37% and the least represented Jura with 4% of total observations. Compare to the population of each canton, the distributions by cantons seems representative of Western Switzerland area. Except for the canton of Geneva which is under-represented and should be around 20%

Variable	**Mean**	**Standard deviation**	**Median**	**Minimum**	**Maximum**
Floorspace	87,38	42,13	83	5	492
Number of rooms	3,51	1,36	3,5	1	15
Population	129 861,82	126 011,14	80 423	1 282	480 627
Employment number	65 473,88	78 660,58	33 392,09	270,41	283 152,85
Household size	2,42	0,67	2,33	1	6
Age of buildings	54,36	26,23	53	3	100
Road noise by night	40,62	9,48	40,02	0	75,33
Train noise by night	13,78	14,92	10,42	0	66,95
Accessibility	15,13	12,72	11,59	0,04	63,23
Canton GE	0,09	0,28	0	0	1
Canton JU	0,04	0,19	0	0	1
Canton NE	0,13	0,33	0	0	1
Canton VD	0,37	0,48	0	0	1
Canton VS	0,19	0,4	0	0	1
Canton FR	0,18	0,39	0	0	1

Table 2: Descriptive statistics

5 Application

This section applies and discusses the methodology presented in Section 3, to real estate announcements data presented in Section 4. The application is done in two steps using the variables presented in Table 1. First, several hedonic regression are applied in order to study the individual contribution of accessibility indicators. Then, based on the BIC criterion, the best model is selected to apply the SEM methodology. The application is done on 3 classes in order to study the distribution of observations in the classes as well as the differences in coefficient estimates.

5.1 Hedonic models

Two accessibility indicators have been defined in Section 3.1. In order to assess their impact on the accuracy of the price estimate, the following 3 models are evaluated and compared :

- Model 1 : without accessibility indicator
- Model 2 : with A^1, accessibility indicator using euclidean distance
- Model 3 : with A^2, accessibility indicator using gravity model

The specifications of accesibility indicators are presented in Section 3.1.

BIC criterion

By definition, the best model is the one that minimizes the BIC. According to the results presented in Table 3, the model with accessibility indicator using generalized cost (Model 3), is better than others. Model with accessibility indicator using distances as the crow flies (Model 2) has a better loglikelihood but the high number of parameters affects its BIC.

This result confirm our intuitions that accesibility indicator improves the accuracy of the price definition and that the difference in altitude has a significant impact.

The values of the coefficients of model with accessibility indicator using generalized cost (Model 3) will be discussed and compared with the SEM approach in Section 5.2. All hedonic regression results are available in Appendix A.

	BIC	log likelihood
Model 1	-525,73	328,78
Model 2	-519,19	343,09
Model 3	**-529,87**	335,24

Table 3: Results of BIC and log likelihood

5.2 Hedonic model vs. SEM algorithm

This section prensents and discusses the main results of the SEM algorithm applied to the model selected in Section 5.1, the spatial distribution of classes, the differences between a OLS method and the SEM algorithm on the signs of coefficients and values, the accuracy and overall price estimation.

SEM algorithm : Main results

First, the main results of the SEM 3-class model and the distribution of the prices of each class in relation to the total distribution are presented in Table 4 and Figure 2. Class 3 is composed of 58% of the observations and has an average price of 22.08 CHF/m^2, located in the centre of the total distribution of prices by the model. Class 1 with a proportion of 37%, an average price of 23.45 CHF/m^2 and a larger standard deviation than class 3 (9,89 vs. 7,59) has prices located in the lower or upper parts of the total distribution. Finally, class 2 with a proportion of 5% completes the first two classes and is more located on lower parts. The weight of this class as well as its standard deviation suggests that it is a collection of outliers.

	Class 1	Class 2	Class 3
Number of observations	2396	358	3806
Weights	0,37	0,05	0,58
Mean	23,45	22,65	22,08
Standard deviation	9,89	16,75	7,59

Table 4: Statistics of the 3-class SEM

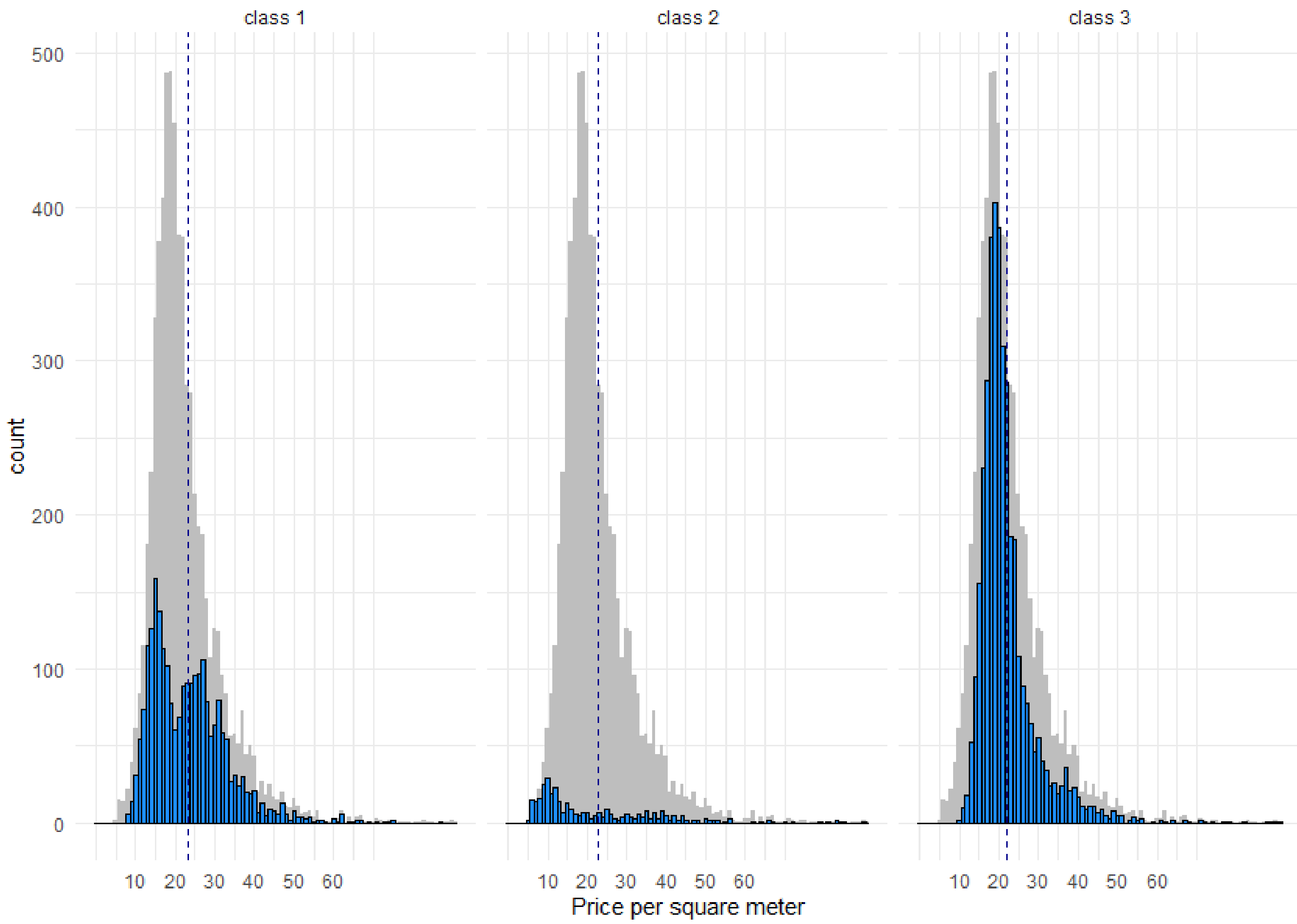

Figure 2: Distribution of classes in the histogram of estimated prices

SEM algorithm : Location of real estate property

Figures 3 to 5 illustrate the spatial distribution of classes 1, 2 & 3. Observations are well distributed in space and not geographically attached to a location (example: centered on the Canton of Vaud). Moreover, prices over floorspace are fairly well distributed and go from less than 16 to more than 28 for all classes. Then, there is not one class that gather observations around one location (canton or city) or absorbs all the rents with the highest (or lowest) prices per square meter.

We note that the dwellings in class 1, with high prices, are particularly concentrated around the Lake Geneva area (Fig. 3). This is not the case for class 3 (Fig. 4), and in particular for dwellings located in the Canton of Geneva, whose prices over floor space appear to be lower than in the other Cantons. The prices of goods located in the countryside seem to be well distributed for each of these two classes. Dwellings of class 2 (Fig. 4), identify as probable outliers, are, despite the small number of observations, well spatially distributed.

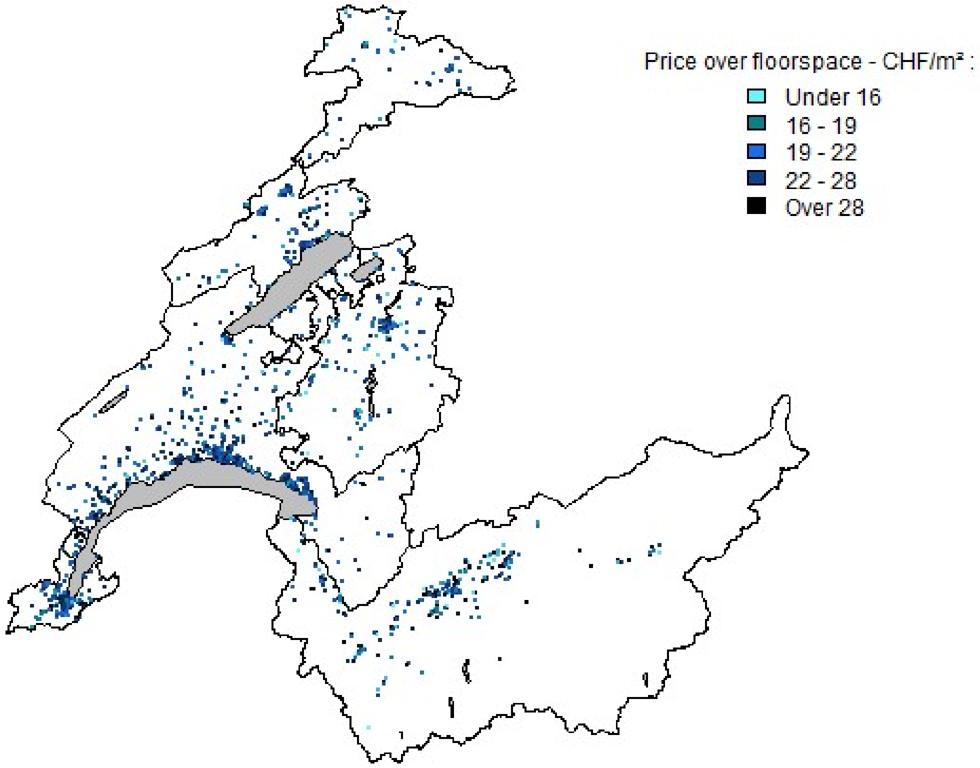

Figure 3: Location of real estate property - class 1

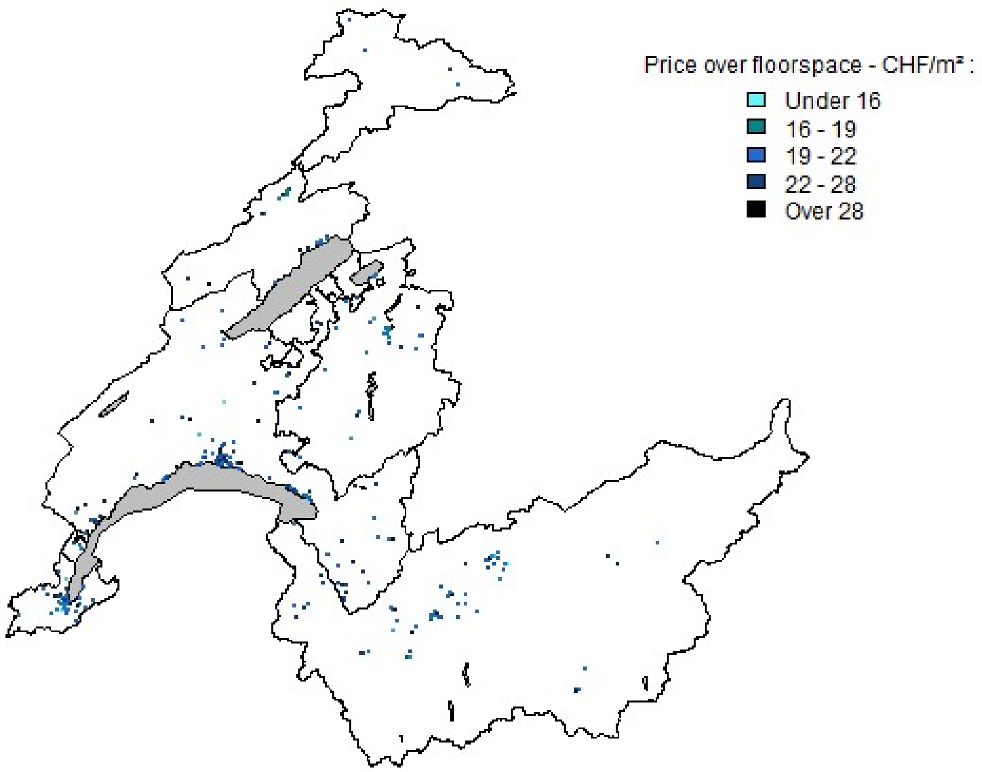

Figure 4: Location of real estate property - class 2

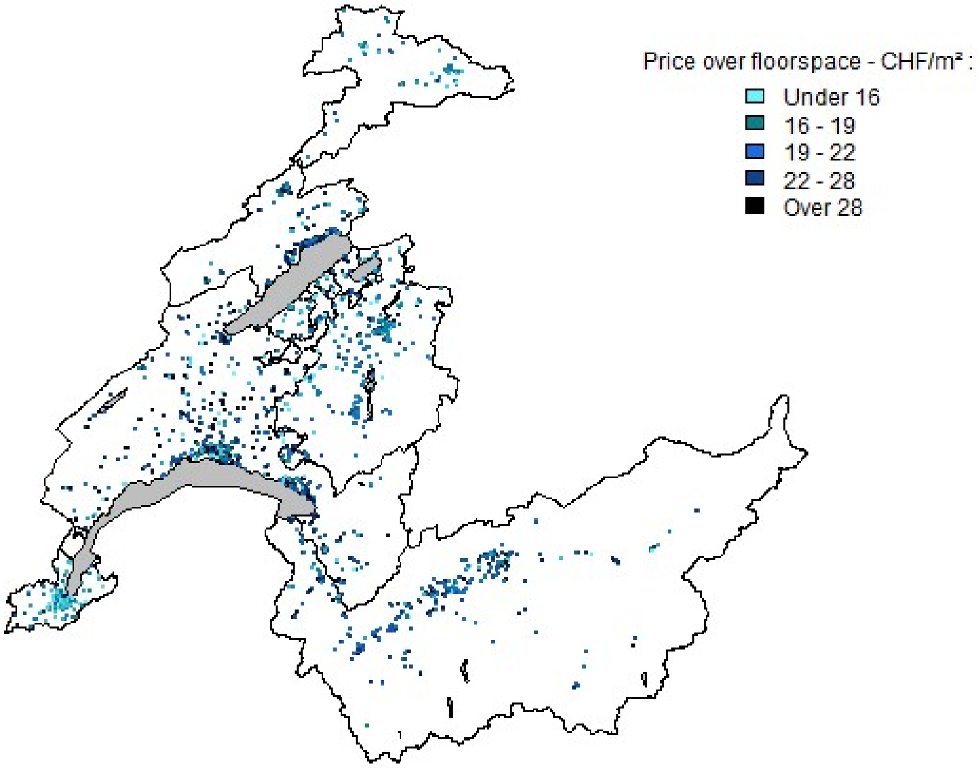

Figure 5: Location of real estate property - class 3

Hedonic model vs SEM algorithm

The, Table 5 presents the results of the coefficients for each of the variables used, the different R^2 and $RMSE$ of the hedonic pricing model and the SEM model.

Let's start with the sign of the variables coefficients of the hedonic model. The signs obtained are all in line with our expectations knowing that the variable explained is the log of the price over floorspace. Even if at first it may seem counter-intuitive, it is consistent that the coefficient of the log of the surface is negative knowing that, generally, the price per square meter decreases when the total surface increases. Following the same reasoning as the floorspace, when the size of the household increases, the number of (bed)rooms usually increases, therefore the floorspace and then, the price over floorspace decreases. Positive coefficients for the number of rooms, the employement number or the population means that an increase in one of these variables implies a price increasement. Which seems intuitive.

The negative coefficient of age of buildings means that the price per square metre decreases as age increases. Explicable given that a new property (or one located in a rewer neighbourhood) is rented for more than the same one in an older neighbourhood. As expected, when road noise increases, the price decreases. But, surprisingly, the coefficient associated with train noise is positive. Only dwellings close to railway stations or rails are affected by train noise.

Moreover, the most important noises are generally near railway stations, where the dwellings are most in demand due to their location. It is therefore possible that train noise is correlated with attractiveness, which would explain the positive sign. The accessibility indicator is positive, indicating that a dwelling close to public transport is more attractive. Finally, the coefficients related to the different cantons also have the expected signs. The positive signs for the cantons of Vaud and Geneva indicate that, for an identical good, the price is higher in these cantons compared to the canton of Fribourg. And vice versa for negative signs. Moreover, if we look at the values, we see that the price is higher for an apartment in Geneva than for an apartment in Vaud.

Let's now study the results of the SEM model, compare to the hedonic model.
It is worth noting that the signs of the significant coefficients are, in most cases, aligned. But, this is not the case in class 1 for the employment number and the household size. It was previously highlighted that dwellings of this class has prices located in the lower or upper parts of the total distribution and especially concentrated around the Lake Geneva area. Several highly speculative explanations are possible. For example, dwellings with high prices over floorspace can be high-end properties belonging to well-to-do households or small apartments for which there is a large supply, much sought-after among students. The majority of students are not concerned by jobs nearby and well-to-do households are not necessarily within a 10-km radius. A more general explanation would be that, given that the Lake Geneva region is a dynamic region with many job opportunities, people do not necessarily need to be close to their jobs within a 10km radius. The negative value of the household size may indicate that these households, with an additional person, prefer to increase the overall quality of the dwelling. This translates into a newer dwelling, better placed or with a nice view. It then increase the price, without increasing the floorspace.
For Class 3, the only significant coefficient different from the hedonic pricing model is train noise. Dwellings of this class account for 58% of the data set. Based on the reasoning developed earlier, this result indicates that the dwellings affected by noise are not attractive in terms of location.
The accessibility indicator is larger for class 3 than for class 2. This result indicates that for the majority of dwellings, being close to public transport increases attractiveness. This is in line with our expectations since Class 3 has a lower price over floorspace. Therefore, on average, households are less well-off, preferring to use public transport, which represents lower expenses compare to the car.
By definition, goods in the same latent class are similar in terms of price definition while goods in different classes are less similar. This, explains the differences of significant variables by class and the weight of the significant coefficients. For example, the variable Canton Jura is significantly different from 0 for class 1 and 3 but not for class 2. At the same time its weight is

lower in the price definition for class 1 than for class 3.

Finally, the performance of the models gives interesting results. We notice that the class 2 of SEM has, as the same time, a R^2 much lower than the other classes and a high $RMSE$. Taking into account the small size of this class (Table 4) this result confirm that class 2 gathers the outliers of the dataset. Class 1 and class 3 have R^2 higher and $RMSE$ lower than the hedonic price model. This means that these 2 classes better describe, for the sub-populations concerned, goods prices. Finally, the overall $RMSE$ of the SEM model is lower than the $RMSE$ of the hedonic price model (0.1535 vs. 0.2109). This result indicates that, overall, the SEM model predicts real estate prices better than the hedonic pricing model.

Variable	Hedonic model			SEM								
				Class 1			Class 2			Class 3		
	Est.	t-stat		Est.	t-stat		Est.	t-stat		Est.	t-stat	
Intercept	3,6487059	85,61	***	3,4219712	92,697	***	3,308495	7,572	***	3,957108	162,19	***
Log Floorspace	-2,0821706	-42,637	***	-2,2092769	-60,631	***	-0,9799	-2,352	*	-2,3124266	-76,728	***
Number of rooms	0,0572323	13,961	***	0,0545457	17,827	***	0,069975	2,013	*	0,0595936	23,621	***
Log Population	0,5759117	4,372	***	1,4454467	11,36	***	-0,23329	-0,17		0,34291	4,924	***
Log Employment number	0,0494881	0,39		-0,7785519	-6,483	***	0,567584	0,421		0,093896	1,392	
Log household size	-0,0016059	-0,08		0,0298574	1,837	.	-0,389512	-1,723	.	0,0090337	0,82	
Log age of buildings	-0,0691474	-4,857	***	-0,0502028	-3,911	***	-0,18752	-0,931		-0,0497994	-6,829	***
Noise by night												
Road	-0,1332282	-3,694	***	0,0559663	1,727	.	-0,924821	-2,817	**	0,0096792	0,478	
Train	0,1422444	5,962	***	0,1810155	9,517	***	0,840619	2,835	**	-0,1115721	-8,286	***
Accessibility indicator	0,0012684	3,592	***	0,0007101	2,638	**	-0,002053	-0,494		0,00368	17,721	***
Canton (ref. is fribourg)												
Geneva	0,2584946	16,42	***	0,1929813	14,939	***	0,428102	2,405	*	0,3912124	42,69	***
Jura	-0,1920417	-10,247	***	-0,2013269	-11,738	***	-0,394403	-1,295		-0,1494007	-15,449	***
Neuchatel	-0,1485638	-12,564	***	-0,1904294	-17,36	***	-0,203258	-1,07		-0,0055005	-0,865	
Vaud	0,1419121	16,883	***	0,3721786	45,795	***	0,211194	1,651	.	0,0138379	3,225	**
Valais	-0,0800155	-8,407	***	-0,1481034	-16,319	***	0,01098	0,08		-0,0108554	-2,221	*
R^2	0.5915			0.9214			0.1464			0.8812		
RMSE	0.2299			0.1113			0.6301			0.0951		
				0.1745								

Table 5: Results of Hedonic model and SEM - 3 classes

Figure 6 completes the results of Table 5 by highlighting the different density functions with respect to the distribution of real price values. As expected, the predicted price density function of the SEM 3-class model is closer to reality than the standard hedonic model. If the differences are always large in the middle of the distributions, we can see that this model is better on high and low values.

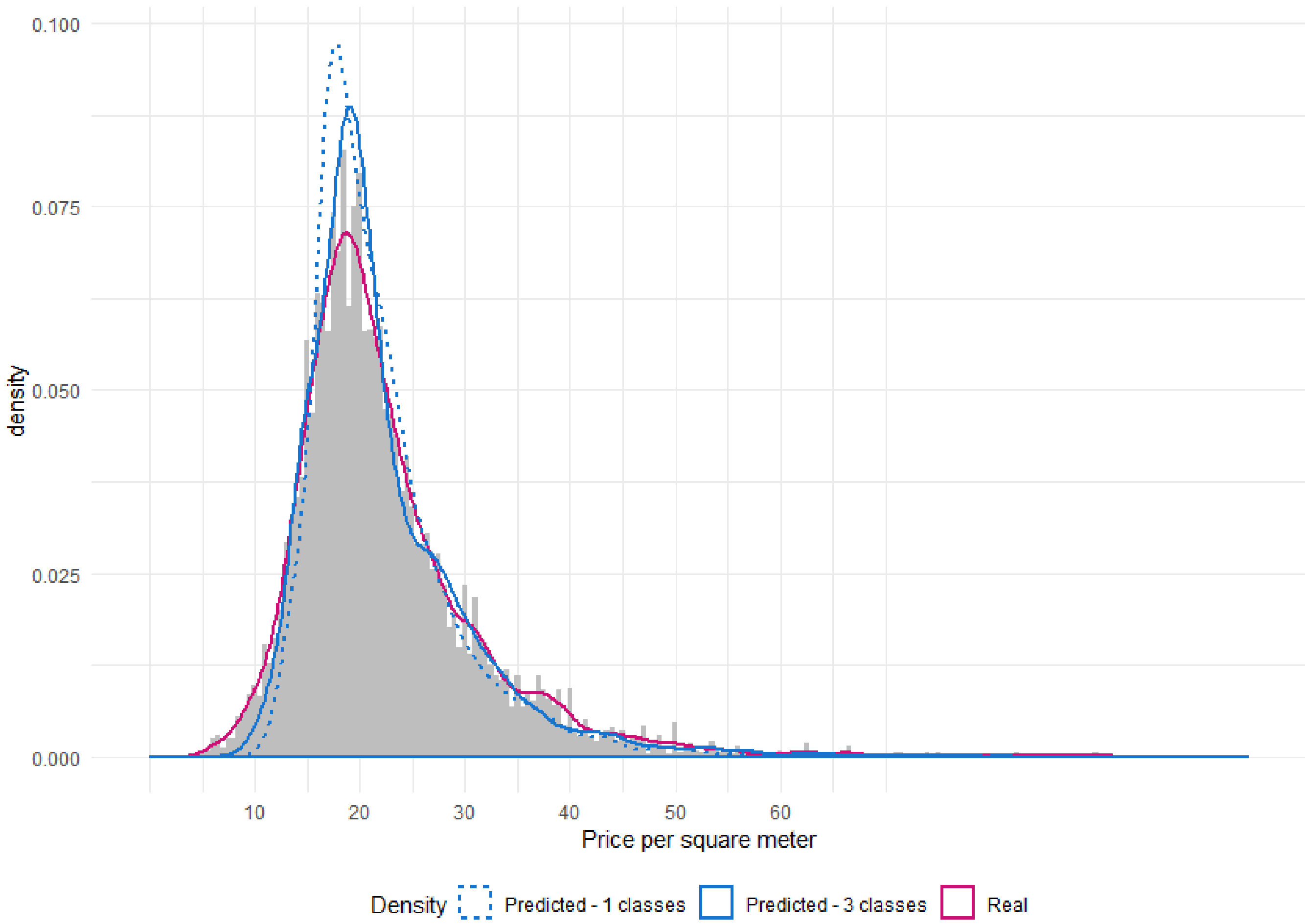

Figure 6: Histogram of real data and density comparison

6 Conclusion

This paper proposes to study the impact of public transport accessibility on the definition of property prices, and more particularly the differences in the context of latent classes (Section 3). Two accessibility indicator are defined : one based on the distance from the dwelling to public transport stops, counting opportunities as the crow flies; the other integrating both the distance and the difference in altitude from the dwelling to public transport stops, using generalized cost, and more specifically the notion of "effort kilometer" used for mountain trails. It presents a methodology combining the Simulated Expectation-Maximization algorithm (SEM) and a criterion measuring the accuracy of statistical models. The SEM algorithm determines the membership of each dwelling to one of the 3 latent classes and the associated coefficients. While the accuracy criterion allows to estimate, in detail, the differences in estimates of each of the latent classes and the overall SEM model compared to the real price of the goods.

The methodology is applied to real estate announcement dataset, built up on combining several sources of open data and containing observations of Western Switzerland (Section 4). First, results highlight that accessibility indicators for public transport improves the accuracy of the price definition and its impact is greater when you consider the difference in elevation. Then, based on the accuracy criterion (RMSE), the SEM algorithm with 3 classes is, overall, more efficient than a hedonic pricing model. In detail, 2 of the 3 classes, representing 95% of the whole dataset, are much more accurate than a hedonic pricing model and one class, representing 5% of the dataset, gathering what should correspond to outliers, is much less accurate. Finally, significant variables are not the same for each class and that the values of the coefficients can vary. These results confirm our initial intuition that there are latent classes for the price definition of dwellings and that they are not evaluated in the same way and therefore on the same characteristics.

The data used for the application are rental announcement data. Even if they give a trend, they do not really represent the real estate market in Western Switzerland. This is why, for future work, we would like to apply this methodology to real transaction data and to study the evolution of the market over time. Then, we wish to develop a methodology capable of predicting the price of a new good by estimating its probability of belonging to a class (supervised clustering). Finally, we will be able to improve our model by improving the performance of the variables used , changing the functional form of the equations or by calculating and comparing hedonic prices of certain characteristics per class and/or over time.

7 References

Baranzini, A., J. Ramirez, C. Schaerer and P. Thalmann (2008) *Hedonic Methods in Housing Markets. Pricing Environmental Amenities and Segregation*, 01 2008.

Baranzini, A., C. Schaerer and P. Thalmann (2010) Using measured instead of perceived noise in hedonic models, *Transportation Research Part D: Transport and Environment*, **15** (8) 473 – 482, ISSN 1361-9209.

Bishop, K. C. and C. Timmins (2011) Hedonic prices and implicit markets: Estimating marginal willingness to pay for differentiated products without instrumental variables, *Working Paper*, **17611**, National Bureau of Economic Research, November 2011.

Boucq, E. and F. Papon (2008) Assessment of the real estate benefits due to accessibility gains brought by a transport project: The impacts of a light rail infrastructure improvement in the hauts-de-seine department, *European Transport Trasporti Europei*, **40**, 51–68, 01 2008.

Cavailhès, J. (2005) Le prix des attributs du logement, ISSN 0336-1454.

Cavailhès, J., T. Brossard, J.-C. Foltête, M. Hilal, D. Joly, F. Tourneux, C. Tritz and P. Wavresky (2009) Gis-based hedonic pricing of landscape, *Environmental and Resource Economics*, **44**, 571–590, 12 2009.

Cebula, R. J. (2009) The Hedonic Pricing Model Applied to the Housing Market of the City of Savannah and Its Savannah Historic Landmark District, *The Review of Regional Studies*, **39** (1) 9–22.

Chen, Y., M. Yazdani, M. Mojtahedi and S. Newton (2019) The impact on neighbourhood residential property valuations of a newly proposed public transport project: The sydney northwest metro case study, *Transportation Research Interdisciplinary Perspectives*, **3**, 100070, 11 2019.

Court, A. (1939) Hedonic price indexes with automotive examples, *The dynamics of automobile demand*, 01 1939.

Dempster, A. P., N. M. Laird and D. B. Rubin (1977) Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society: Series B*, **39**, 1–38.

El-Geneidy, A., P. TÃ©treault and J. Surprenant-Legault (2009) Pedestrian access to transit: Identifying redundancies and gaps using a variable service area analysis, 08 2009.

Ellickson, B. (1981) An alternative test of the hedonic theory of housing markets, *Journal of Urban Economics*, **9** (1) 56–79.

Fraley, C. and A. E. Raftery (2002) Model-based clustering, discriminant analysis, and density estimation, *Journal of the American Statistical Association*, **97** (458) 611–631.

Geurs, K. and B. van Wee (2004) Accessibility evaluation of land-use and transport strategies: Review and research directions, *Elsevier*.

Gorshenin, A. K., V. Y. Korolev and A. M. Tursunbaev (2017) Median modifications of the em-algorithm for separation of mixtures of probability distributions and their applications to the decomposition of volatility of financial indexes, *Journal of Mathematical Sciences*, **227** (2) 176–195, Nov 2017, ISSN 1573-8795.

Griliches, Z. (1961) Hedonic Price Indexes for Automobiles: An Econometric of Quality Change, in *The Price Statistics of the Federal Goverment*, NBER Chapters, 173–196, National Bureau of Economic Research, Inc.

Gross, D. J., C. F. Sirmans and J. D. Benjamin (1990) An empirical evaluation of the probabilistic bid-rent model : The case of homogenous households, *Regional Science and Urban Economics*, **20** (1) 103–110, June 1990.

Grun, B. and F. Leisch (2008) Flexmix version 2: Finite mixtures with concomitant variables and varying and constant parameters, *Journal of Statistical Software, Articles*, **28** (4) 1–35, ISSN 1548-7660.

Han, M. and Y. Xue (2015) Application of expectation maximization algorithm in magnetic induction tomography, *Biomedical Engineering Letters*, **5** (3) 221–228, Sep 2015, ISSN 2093-985X.

Hansen, W. G. (1959) How accessibility shapes land use, *Journal of the American Institute of Planners*, **25** (2) 73–76, May 1959, ISSN 0002-8991.

Hayrullahoglu, G., Y. Aliefendioglu, H. Tanrivermis and A. Hayrullahoglu (2018) Estimation of the hedonic valuation model in housing markets: The case of cukurambar region in cankaya district of ankara province, *Ecoforum*, **7**, 01 2018.

Howell, J. F. and S. Peristiani (1987) The estimation of a hedonic asking and offer rent equation model: An em algorithm approach, *Empirical Economics*, **12** (4) 203–220, Dec 1987, ISSN 1435-8921.

Klimenko, Y. and M. de Lapparent (2019) Modeling housing rental prices in western switzerland, 01 2019.

Kryvobokov, M. and L. Bouzouina (2014) Willingness to pay for accessibility under the conditions of residential segregation.

Laferrere, A. (2005) Hedonic housing price indexes: the French experience, in B. for International Settlements (ed.) *Real estate indicators and financial stability*, vol. 21 of *BIS Papers chapters*, 271–287, Bank for International Settlements, December 2005.

Lazarsfeld, P. F. (1950) The logical and mathematical foundation of latent structure analysis, *Studies in Social Psychology in World War II Vol. IV : Measurement and Prediction*, 362–412.

Limsombunchai, V., C. Gan and M. Lee (2004) House price prediction: Hedonic price model vs. artificial neural network, *American Journal of Applied Sciences*, **1**, 03 2004.

Liu, Z., J. Almhana, V. Choulakian and R. McGorman (2006) Online em algorithm for mixture with application to internet traffic modeling, *Computational Statistics & Data Analysis*, **50** (4) 1052 – 1071, ISSN 0167-9473.

Löchl, M. and K. W. Axhausen (2010) Modelling hedonic residential rents for land use and transport simulation while considering spatial effects, *Journal of Transport and Land Use*, **3** (2), Sep. 2010.

Martinez, L. and J. Viegas (2009) Effects of transportation accessibility on residential property values: Hedonic price model in lisbon, portugal, metropolitan area, *Transportation Research Record*, **2115**, 01 2009.

Mayer, M., S. Bourassa, M. Hoesli and D. Scognamiglio (2018) Estimation and updating methods for hedonic valuation, *SSRN Electronic Journal*, 01 2018.

Melichar, J., P. Rieger and K. Jedlička (2019) Application of hedonic price model in the prague property market, 11 2019.

Moroz, J. R. C. B.-G. X. L. L. B. H. (2017) *Improving the Resilience of Peru's Road Network to Climate Events*, The World Bank.

Mustafaa, Y., V. Tolpekin and A. Stein (2011) Application of the em-algorithm for bayesian network modelling to improve forest growth estimates, *Procedia Environmental Sciences*, **7**, 74 – 79, ISSN 1878-0296. Spatial Statistics 2011: Mapping Global Change.

Oladunni, T. and S. Sharma (2016) Hedonic housing theory : A machine learning investigation, 12 2016.

Picard, N., C. Maurin and A. de Palma (2012) Regime switching models: An application to the real estate market in ile-de-france, 01 2012.

Rosen, S. (1974) Hedonic prices and implicit markets: Product differentiation in pure competition, *Journal of Political Economy*, **82** (1) 34–55.

Ruf, D. (2017) Agglomeration Effects and Liquidity Gradients in Local Rental Housing Markets, *Working Papers on Finance*, **1702**, University of St. Gallen, School of Finance, February 2017.

Song, S. (1996) Some tests of alternative accessibility measures: A population density approach, *Land Economics*, **72** (4) 474–482.

Vickerman, R. W. (1974) Accessibility, attraction, and potential: A review of some concepts and their use in determining mobility, *Environment and Planning A: Economy and Space*, **6** (6) 675–691.

Waugh, V., Frederick (1929) *Quality as a Determinant of Vegetable Prices: a Statistical Study of Quality Factors Influencing Vegetable Prices in the Boston Wholesale Market*, New York.

Wezel, M., M. Kagie and R. Potharst (2005) Boosting the accuracy of hedonic pricing models, *Erasmus University Rotterdam, Econometric Institute, Econometric Institute Report*, 12 2005.

Wing, C. K. and T. Chin (2003) A critical review of literature on the hedonic price model, *International Journal for Housing Science and Its Applications*, **27**, 145–165, 06 2003.

Yang, Y., I. M. Longini Jr., M. E. Halloran and V. Obenchain (2012) A hybrid em and monte carlo em algorithm and its application to analysis of transmission of infectious diseases, *Biometrics*, **68** (4) 1238–1249.

A Hedonic models

OLS - Model 1

Model 1 is composed of the variables presented in Table 1, without the public transport accessibility variable.

	Est.	Std. Error	t value	Pr(>|t|)	
Intercept	3,609968	0,041271	87,47	< 2e-16	***
Log Floorspace	-2,080122	0,048876	-42,559	< 2e-16	***
Number of rooms	0,05629	0,004095	13,746	< 2e-16	***
Log Population	0,529663	0,131223	4,036	0,0000549	***
Log Employment number	0,141967	0,124358	1,142	0,253663	
Log household size	-0,017129	0,019543	-0,876	0,380792	
Log age of buildings	-0,063467	0,014163	-4,481	0,00000754	***
Noise by night					
Road	-0,125697	0,036034	-3,488	0,000489	***
Train	0,171803	0,022415	7,665	2,05E-14	***
Canton (ref. is fribourg)					
Geneva	0,276541	0,014933	18,519	< 2e-16	***
Jura	-0,18935	0,018743	-10,103	< 2e-16	***
Neuchatel	-0,147498	0,011831	-12,467	< 2e-16	***
Vaud	0,140734	0,008407	16,74	< 2e-16	***
Valais	-0,081719	0,009515	-8,588	< 2e-16	***
R^2	0,5907				

Table 6: OLS - Model 1

OLS - Model 2

Model 2 is composed of the variables presented in Table 1, including a public transport accessibility variable as :

$$A_i^1 = \sum N_j \; I(d_{i,j} < \bar{d}) \quad (15)$$

With N_j the number of opportunities at location j , $d_{i,j}$ the distance, in meters, between the dwelling and the bus stop, the subway or the train station and $\bar{d}$ a specific radius (200 m / 400 m / 600 m and 800 m).

	Est.	Std. Error	t value	Pr(>|t|)	
Intercept	3,6447446	0,0421396	86,492	< 2e-16	***
Log Floorspace	-2,0858544	0,0488119	-42,732	< 2e-16	***
Number of rooms	0,0574209	0,004097	14,015	< 2e-16	***
Log Population	0,567494	0,1314494	4,317	0,000016	***
Log Employment number	0,0624061	0,1256913	0,497	0,619556	
Log household size	-0,0031088	0,0199623	-0,156	0,876248	
Log age of buildings	-0,0679554	0,0144556	-4,701	0,00000264	***
Noise by night					
Road	-0,1194956	0,0361922	-3,302	0,000966	***
Train	0,1353664	0,0237217	5,706	0,000000012	***
Accessibility indicator					
< 200m	-0,006069	0,0025447	-2,385	0,017112	*
200m - 400m	0,0018693	0,00151	1,238	0,215776	
400m - 600m	0,0040228	0,0013344	3,015	0,002582	**
600m - 800m	0,0006176	0,0010833	0,57	0,568619	
Canton (ref. is fribourg)					
Geneva	0,2695398	0,0149687	18,007	< 2e-16	***
Jura	-0,1952961	0,0187576	-10,412	< 2e-16	***
Neuchatel	-0,1542372	0,0119271	-12,932	< 2e-16	***
Vaud	0,1403589	0,0084008	16,708	< 2e-16	***
Valais	-0,0842426	0,0095201	-8,849	< 2e-16	***
R^2	0,5925				

Table 7: OLS - Model 2

OLS - Model 3 Model 3 is composed of the variables presented in Table 1, including a public transport accessibility variable as :

$$A_i^2 = \sum N_j \exp^{\lambda(\frac{d_{i,j}}{1'000} + \frac{\Delta h_{i,j}}{100})} \tag{16}$$

With, N_j the number of opportunities at location j, $d_{i,j}$ the distance in meters, $\Delta h_{i,j}$ the elevation difference in meters and λ a parameter fixed to -1.

	Est.	Std. Error	t value	Pr(>|t|)	
Intercept	3,6487059	0,0426202	85,61	< 2e-16	***
Log Floorspace	-2,0821706	0,0488349	-42,637	< 2e-16	***
Number of rooms	0,0572323	0,0040996	13,961	< 2e-16	***
Log Population	0,5759117	0,1317343	4,372	0,0000125	***
Log Employment number	0,0494881	0,1268839	0,39	0,696529	
Log household size	-0,0016059	0,0199979	-0,08	0,935998	
Log age of buildings	-0,0691474	0,0142377	-4,857	0,00000122	***
Noise by night					
Road	-0,1332282	0,0360623	-3,694	0,000222	***
Train	0,1422444	0,0238582	5,962	2,62E-09	***
Accessibility indicator	0,0012684	0,0003531	3,592	0,00033	***
Canton (ref. is fribourg)					
Geneva	0,2584946	0,0157423	16,42	< 2e-16	***
Jura	-0,1920417	0,0187405	-10,247	< 2e-16	***
Neuchatel	-0,1485638	0,0118243	-12,564	< 2e-16	***
Vaud	0,1419121	0,0084057	16,883	< 2e-16	***
Valais	-0,0800155	0,0095181	-8,407	< 2e-16	***
R^2	0,5915				

Table 8: OLS - Model 3