CHAPTER 2
Index Construction for Beginners

Since housing market indices play an important role throughout this entire thesis, this chapter provides a brief overview of the applied index construction methods. To study the performance of asset markets, good quality indices are a first requirement. This sounds reasonable, but for real estate market, constructing such indices can be cumbersome. Every stock market is covered by various indices, based on market segments or the entire local listed share universe. These indices reflect actual market conditions many times per day. In contrast, a real estate market index cannot be based on the entire universe of properties, since properties are traded very infrequently. These properties do change in value over time without actual trading, and are very heterogeneous. When constructing a real estate index by means of transaction prices, the package of sold properties is assumed to be a representative sample of the property stock, and the time interval in between index numbers has to be rather large in order to avoid unrealistic volatility. Using appraised values instead of transaction prices does not help much, since actual appraisals occur once in a number of years. Moreover, appraisals often do not reflect true market value. Obtaining a reliable real estate market index is not as straightforward as it might seem at first sight, and is far more difficult than obtaining a reliable stock market index.

This is a problem, since investors need to judge the attractiveness of real estate as an asset class, compared with investment assets like stocks or bonds. This attractiveness of the real estate market depends on its returns and risks, which can only be measured by means of a reliable index. For example, due to the use of less appropriate indices, many studies show unrealistically low standard deviation of real estate returns, such that a mean-variance portfolio optimizer might erroneously allocate a large part of a mixed asset portfolio to real estate. Examples of studies showing very low standard deviations of real estate returns relative to other assets are Ross and Zisler (1991), Goetzmann and Ibbotson (1990) and Gilliberto (1990), and are reviewed by Norman, Sirmans and Benjamin (1995). Without a correct index, the real estate market cannot be analyzed accurately with respect to attractiveness, behavior, efficiency, anomalies or other aspects commonly studied for stock or bond markets. A reliable index is therefore very desirable.

Transaction prices
Residential real estate usually provides sufficient numbers of transactions for reliable index construction, in contrast with commercial real estate like offices, retail and industrial properties. This is caused by the fact that the total amount of commercial
properties is substantially lower than the amount of residential properties, and by the fact that homogeneity is much stronger for residential property than for commercial property. The consensus is that an index based on mean or median transaction prices is not very accurate, due to violation of the sample representativeness assumption. The distribution of house prices is skewed to the right, which means that an index based on mean prices is very sensitive to the exact number of high-price properties sold. For example, if villas are sold more frequently during the spring than during other seasons, the mean index will show a peak for every spring. This index does not only reflect market changes, but also quality changes in the sample composition. In other words, this index is not a constant-quality index. A median index is less sensitive to this problem, but is still far from perfect. Other index construction methods have been developed and applied in practice to account for this changing sample quality.

**Appraised values**

As we stated before, the absolute number of transactions for commercial real estate is lower, such that a transaction-based index will generally not be viable in practice. Therefore, indices reflecting commercial real estate class are mostly based on appraised values. Geltner (1991, 1993a) showed how appraisal based values cause smoothing and lagging if used to construct a real estate index, thereby altering both index levels and volatility. Anchoring on previous appraisals and conservatism of the appraiser are two major sources of these distortions. Besides techniques to adjust volatility of an appraisal-based real estate series and to remove the lag, Geltner and others applied constant-quality index techniques to transaction prices of commercial real estate, which have originally been applied to house prices (Geltner (1996), Fisher and Geltner (2001), Gatzlaff and Geltner (1998), and Geltner and Goetzmann (2000)).

**Constant quality**

Two ‘rival’ constant-quality index methods are hedonic indices and repeat-sales indices. These techniques correct for possible index distortions caused by changing sample quality, but in different ways. Hedonic index methods explicitly correct for property quality differences by homogenizing properties and property prices using a set of quality variables to filter price observations. Repeat-sales methods do not look at price levels, and use price changes of the same property instead. Wallace (1996) suggests that in case of large numbers of repeat transactions and constant housing attribute levels and prices, the hedonic-based indices and repeat-sales indices are likely to be equivalent, in theory. Sometimes a mixture of both techniques is used, which is referred to as a hybrid or a hedonic repeated-measures method. Excellent studies have been published reviewing constant quality index construction, like Cho (1996) and Wang and Zorn (1997), so we do not intend to provide another literature review. However, since repeat-sales techniques and hedonic techniques are used throughout all following chapters, we discuss methodological issues of these techniques in more detail now.
2.1 Hedonic indices

A house could be perceived as a composite of attributes. The house price could therefore be considered as a function of implicit prices for those attributes, plus a premium or a discount determined by general market conditions. For example, if interest rates fall, all house prices are likely to change, while implicit attribute prices will remain constant. A hedonic regression tries to reveal the hedonic price function for property attributes, while the remaining unexplained part of house prices is assumed to represent 'the market'. These market prices are transformed into index numbers. In practice, hedonic indices are widely used; for example, the national house price indices of Halifax and Nationwide in the United Kingdom are constructed by means of hedonic regression.

After the first application by Court (1939) on car prices, the hedonic methodology was theoretically founded by Tinbergen (1959) and Rosen (1974). Besides creating real estate indices, hedonic regression is also used to determine housing market externalities, as discussed in Chapter 6. The hedonic theory is based on the idea that consumers try to optimize consumption of housing attributes by selecting the bundle of attributes, the property, which is closest to the preferred one, depending on the implicit attribute prices and the available budget (Kokoski et al, 1999). Derivation of the implicit prices, which assumes market equilibrium, occurs through estimation of a regression equation like (2.1).

\[
 f(P_{it}) = \sum_{t=0}^{T} \beta_i D_{it} + \sum_{k=1}^{N} a_k g(X_{ki}) + \epsilon_{it},
\]

\(P_{it}\) denotes the price of property \(i\) sold at time \(t\), \(X_i\) represents the vector containing physical and locational attributes of this property, and dummy \(D_t\) specifies whether the property is sold at time period \(t\). The estimate for \(\beta_i\) can then be simply transformed into index number \(I_i\) by the transformation:

\[
 I_i = 100 \frac{f^{-1}(\hat{\beta}_i)}{f^{-1}(\hat{\beta}_0)}
\]

In Equation (2.1), a transformation of the property price is regressed on time dummies and a function of the property characteristics. Box-Cox transformations can be used to test for the empirically best transformation of the variables, since there is no best specification a priori. Examples of studies dealing with specification issues are Burgess and Harmon (1991), Anglin and Genkay (1996) and Pace (1993). Many studies find the semi-log specification to be the most convenient, in which the natural logarithm of house prices is expressed as a linear function of attributes and time dummies. In that case, prices of attributes and time dummies are estimated as log price changes, instead of price levels.
The exact specification of the hedonic regression equation could take many forms. If the actual price is regressed on actual characteristics, Equation (2.1) assumes that implicit attribute prices remain constant throughout the entire sample period. To relax this strict assumption, a semi-log specification allows for varying attribute prices over time, since these prices are estimated as a percentage of the sales price instead of an absolute amount. For example, if the presence of a garage is valued at 10 percent of the sales price, this value could be €15,000 in 2000 and €16,500 in 2001. In that case, percentage attribute prices are assumed to be constant. An alternative specification, like Equation (2.3), could allow explicitly for time varying attribute prices.

\[
(2.3) \quad f(P_{t,j}) = \sum_{t=t_0}^{T} D_t \left( \beta_t + \sum_{k=1}^{N} a_{k,t} g(X_{k,i}) \right) + \epsilon_{t,i}
\]

The advantage of the hedonic approach relative to its other constant-quality index ‘rival’ is the avoidance of biases in regression outcomes due to the specific calculation and wasteful data selection procedures, as discussed in the next section. However, the ‘correct’ attribute prices can only be estimated if the ‘correct’ functional form of the regression equation is chosen, and if the ‘correct’ set of attributes is selected. Theory does not provide guidelines, and this potential misspecification of the functional form and the possibly incorrect variable selection are two major impediments to the reliability of hedonic regression outcomes. Moreover, econometric problems like multicollinearity might disturb estimates. In addition, although probably less severe than with repeat-sales estimation, sample selection bias can still not be excluded a priori, as Zuehlke (1989) and Munneke and Slade (2000) show.

More recently, spatial estimation techniques are used to mitigate the inefficiency caused by omitted variables, and to allow for space varying implicit attribute prices. Chapter 6 provides a literature discussion of this technique, which will be applied in Chapters 7 and 8.

Although frequently applied to housing, hedonic regression has been applied to construct commercial real estate index series as well. Examples are Saderion, Smith and Smith (1994) and Munneke and Slade (2000).

However, limited or costly data availability might prevent one from estimating hedonic real estate index series, especially for very broad market coverage or very long time series. In these cases, it will often be impossible to obtain a complete set of quality variables. This is a very important reason to apply a repeat-sales regression instead of the hedonic methodology.
2.2 Repeat-sales indices

A repeat-sales index is based on the price difference between repeated transactions of the same asset. If a property is sold at least twice within the sample period, and if it did not change in quality, the log price difference between two succeeding transactions is regressed on dummy variables. Since only the transaction prices and transaction moments are used for index estimation, a repeat-sales index does not suffer from the specification and variable selection problems encumbering the hedonic regression technique. In practice, the repeat-sales methodology is widely used as well. For example, the U.S. federal government (Office of Federal Housing Enterprise Oversight, OFHEO) constructs house price indices by means of so-called geometric weighted repeat-sales regression.

The repeat-sales method was first proposed in the seminal study of Bailey, Muth and Nourse (1963). In contrast with hedonic regression, the functional form of the regression equation is not prone to discussion, except for potential inclusion of a constant. In almost all repeat-sales studies, logarithmic price differences are used. Bailey, Muth and Nourse used OLS to estimate Equation (2.4).

\[
\ln \left( \frac{P_{t,i}}{P_{t-1,i}} \right) = \sum_{t=t_0}^{t_1} \beta_t D_{t,i} + \epsilon_{t,i}
\]

Estimates of parameters \( \beta_t \) are transformed into index numbers \( I_t \) by

\[
I_t = 100 \left( \frac{\exp(\beta_t)}{\exp(\beta_{t_0})} \right)
\]

Time indicators \( t_1 \) and \( t_2 \) denote the moments of the first and succeeding sale, respectively, and \( P_t \) the respective prices. The log price differences are regressed on dummy variables \( D_t \) which could be specified in a geometric or an arithmetic manner. In the geometric specification, dummies are zero, except for the moments of first and second sale, where dummies are -1 and +1, respectively. In the arithmetic specification, all dummies \( D_{t+1,i} \) through \( D_{t,2} \) are one, and zero otherwise. Wang and Zorn (1997) argue that the arithmetic specification is the most appropriate alternative in a housing index context. Goetzmann (1992) suggests an ex-post adjustment to derive expected values of arithmetic means by

\[
I_t = 100 \exp \left( \hat{\beta}_t + \frac{1}{2} \hat{\sigma}_t^2 \right)
\]

The repeat-sales technology became popular since Case and Shiller (1989) modified the method of Bailey, Muth and Nourse. Case and Shiller showed that the variance of OLS residuals from Equation (2.4) rises with the time interval between sales, and proposed a weighted least squares approach to correct for this specific heteroskedasticity. Since the
original observations will be weighted with respect to the OLS regression, the index resulting from the Case-Shiller estimation is called a weighted repeat-sales index. This is the most common approach to repeat-sales index construction.

The repeat-sales methodology received a lot of attention in academic publications, mostly concerned with biases arising from the specific calculation method and the data selection. One major source of bias results from the potential non-representativeness of the property sample. By definition, for any transaction-based index, only transacted properties are selected for index construction. Repeat-sales indices, however, only select properties that are sold at least twice. As a result, frequently transacted properties like 'starter homes' will be overrepresented in the sample. Also properties bought because of speculation motives will sell after shorter holding periods with larger price increases, thereby distorting estimates. The likely resulting bias is studied by Case, Pollakowski and Wachter (1997), Gatzlaff and Haurin (1997) and Clapp and Giaccotto (1992b), amongst others.

Repeat-sales indices generally have low frequency of observations relative to stock indices. This is caused by infrequent trading, but it poses a potential threat to the reliability of the regression estimates, arising from the temporal aggregation of observations. If the chosen time interval deviates from the 'correct' interval, the calculated index might erroneously include or exclude time information. Examples of studies concerning this bias are Englund, Quigley and Redfearn (1999), Geltner (1993b), and Dombrow, Knight and Sirmans (1997).

Other distortions of regression estimates are likely if property attributes change over time, or if the property has been renovated in between two succeeding transactions. Amongst others, other research issues regarding repeat-sales indices concern the revision of previously estimated indices due to the availability of new information later in time (Clapp and Giaccotto, 1998), inclusion of a constant in the regression equation (Goetzmann and Spiegel (1995), Steele and Goy (1997), and Kuo (1997)), and methods for reliable estimation of separate sub-market indices (Goetzmann and Spiegel (1997), and Zabel (1999)).

2.3 Hybrid indices

The advantages of the repeat-sales index are the disadvantages of the hedonic approach, and vice versa. Both methods have desirable characteristics. Therefore, some comparative studies advocate a mixture of both approaches, like Quigley (1995), Clapp and Giaccotto (1998), Meese and Wallace (1997), and Case and Szymanoski (1995).
2.4 Summary and conclusions

The abundance of literature has made clear that hedonic and repeat-sales indices are theoretically superior to simple mean or median indices. In practice, the decision to opt for one specific constant-quality index method is likely to depend on data availability. If the constructed index will be used for performance measurement of portfolio managers, ex post adjustments of previously calculated indices might hamper a repeat-sales index from being used as a benchmark.

Both hedonic and repeat-sales approaches are used throughout this entire thesis. In Chapters 3 and 4, the repeated measures methodology is applied to construct the indices used. Since these index series concern hundreds of years, the motivation for selecting a repeat-sales approach will be obvious. In Chapters 6 through 8, the hedonic method is applied. However, in those chapters, the aim is not to construct a real estate index, but to study housing market externalities. Hedonic regression can be very useful for this purpose. Now the theoretical founding for these approaches has been outlined, it is really time to start using them.
original observations will be weighted with respect to the OLS approach, the index resulting from the Case-Stahle estimation is called a weighted repeat-sales index. This is because the main objective is to reduce the bias of the observed differences in housing prices to provide insight into the nature of the price change in a single market over time. This approach involves using a statistical model to estimate the price change for individual properties in the market, taking into account various factors that influence the price change. These factors include property characteristics, location, and time-specific factors. By applying this approach, the weighted repeat-sales index provides a more accurate measure of the change in housing prices.

As a result, research has shown that the new approach is more effective in capturing the true nature of the price change in a single market. The results also demonstrate that the new approach can be used to identify and quantify the effects of various factors on housing prices. This is an important contribution to the field of real estate research and can be used to inform policy decisions and investment strategies. The data from these studies can provide valuable insights for both policymakers and investors. This is achieved by introducing a technique that reduces the bias of the observed differences in housing prices to provide insight into the nature of the price change in a single market over time. This approach involves using a statistical model to estimate the price change for individual properties in the market, taking into account various factors that influence the price change. These factors include property characteristics, location, and time-specific factors. By applying this approach, the weighted repeat-sales index provides a more accurate measure of the change in housing prices.

2.3 Hybrid indices

The advantage of the new approach is that it can take into account the disadvantages of the hedonic approach, and vice versa. Both methods have desirable characteristics. Therefore, some comparative studies advocate a mixture of both approaches, like Croom (1995), Clapp and Lazear (1997), and Cao and Seyfried (1997)