Abstract: Presented here is a model objectivizing real estate prices so that prices across time could be compared to understand historical price trends and also to assist in a property evaluation or appraisal, as well as for the analysis of comparables in estimating a reasonable offer for a property on the market. Given a timespan of interest, a locale (e.g., a particular zipcode, a city, a county, a state), a category of properties of interest (e.g., condos), an objective historical trend in values can be computed by first evaluating the ratios between the transactions’ realized prices and objective governmental assessment of the properties at some fixed point of time; then, for each period (a month) averaging the ratios of all transaction in that period; then, comparing said averages (or medians) between different periods.

Index Terms: Automatic Valuation Model, Geospatial Data Trend Analysis, House Price Trend Analysis, Real Estate, Spatiotemporal Extrapolation, Spatiotemporal Interpolation, Spatiotemporal Summarization

1. BACKGROUND

Various services and methods exist for the estimation of the change over time in real estate prices in any given locale. Said prior models typically compute the average or median sale price in the locale during each period and then compare said statistics between the various periods. Some of said prior models can also focus their comparison on specific property categories, e.g., single-family homes or condos, and may further narrow the categories down, e.g., 3-bedroom homes or houses of 2000-2500 interior square feet. Yet, in said prior models, there is, in fact, a comparison of apples to oranges. Even in a small locale, e.g., a zipcode, and even in a narrow category, there are vastly different properties being averaged. This creates a statistical bias when different periods are compared since in one period there could dominate sales of quality-built properties with a view, while in another period, lesser properties could dominate. This bias becomes even stronger when larger areas are analyzed, e.g., at the county or state level, because demographic changes can favor sale activity more in cheaper subareas in one period and in more exclusive subareas in another period.

Models exist comparing price per unit of size, e.g., price per interior square foot of a home. However, that too comingles residences with a view and residences without a view, well-built houses to poorly built; further accounting for one size metric, such as interior area, ignores other size metrics, such as the lot size.

A recent improvement to Automatic Valuation Models (AVM) [1-4] of properties includes the computation of ratios of actual sale prices to government-assessed values and the extrapolation of such ratios for the valuation of a specific property.

2. THE PRESENT METHOD

Presented here is a model objectivizing real estate prices so that prices across time can be compared to understand historical price trends and also to assist in property evaluation or appraisal, as well as for the analysis of comparables in estimating a reasonable offer for a property on the market.

In order to objectivize and normalize real estate transactions across a locale and a time period, we need to have a metric of valuation of properties that was consistent among all the properties in the locale at some point in time. Said point in time of the objective metric does not need to be within said period. Further, said metric does not have to represent the true value...
of each property at said point in time; rather, it has to be consistently related or proportional to the true value. Said relationship does not have to be a precise linear proportion, nor does it have to be truly consistent in 100% of the cases since we only need that metric for a statistical aggregation of large numbers of cases. A good candidate for said metric is property valuation by local government tax assessors, particularly the tax appraisal offices in most counties in the United States. Said offices typically invest immense effort in the attempt of consistent valuation of all the properties under their jurisdiction, taking into account quantitative metrics (such as the size of the interior, the size of the lot, year built, year renovated, the ground elevation, the floor level elevation of a condo in a building, the costs of improvement made based on the permits filed, etc.) and qualitative metrics (location, exposure, view, special features, etc.). For example, in Florida, the county assessor offices determine what they call the “just value” of the properties as of January 1 of the assessment year. In order to minimize litigation, the assessor’s office typically sets the “just value” at 10-20% below the true value, which does not affect the algorithm presented here as long as said discount is reasonably consistent.
It should be noted that government offices sometimes provide multiple types of valuations for tax purposes. The following example shows the various official “valuations” available from Florida counties. Among these valuations, the only meaningful one for the present purposes is the “Just Value.” The other valuations either reflect only a part of the property value (e.g., the Land Value and the Improvement Value) affected by the demographics of the property owner and, thus, are not meaningful for understanding the true value of the property.
The method proposed herein compares the transactional sale price of each property, no matter when, to one time-fixed metric of an objective valuation in order to evaluate the ratio by which the realized price is above (or below) said metric. That is, this ratio is the ratio between the realized price and said objective metric. In the example of this figure, sales at different times are compared to the county’s “Just Value” as of January 1, 2021, to compute the Ratio factor. Notice that Row 3 in the table contains an obvious data entry error. Therefore, there can be a data-cleansing process in order to disregard outliers that are outside a reasonable range. Data about the realized prices of each transaction can be obtained from proprietary databases, such as those provided by data consolidators, from county or state records, or from the Real Estate Multiple Listing Service (MLS), as in the following figure.
Figure 3: Ratios of the realized price, at various times, to the County “Just Value” of 2021-0-01. Row 3 is an outlier to be disregarded.

The Ratio thusly computed is an objective comparison metric between different sale transactions in a locale at close times or across long timespans.

To better compare sale transactions over time within a locale, we can subdivide properties into categories because it is possible that in different property categories, prices increased at different paces. For example, we can consider two
categories of properties: single-family homes vs. condominium apartments.

Next, we consider a locale of interest, e.g., Zipcode 33175; a category of interest, e.g., Houses (single-family homes); and a timespan of interest, e.g., from January 1, 2006, through December 31, 2007. We subdivide said timespan into periods, e.g., calendar months. In each period, for each sale transaction, we evaluate the Ratio of the price to the fixed objective metric, e.g., the 2021 County “Just Value.” We can exclude outlier transactions based on any criteria of outlier exclusion. For each period, we evaluate a representative statistical aggregator of the ratios, e.g., the average of the ratios or the median of the ratios, of all the relevant sale transactions. We can further exclude months with a very low number of transactions, e.g., less than 6, to avoid the possibility of excess weight of any single transaction, which may cause bias in statistical analysis across time.

<table>
<thead>
<tr>
<th>Zipcode</th>
<th>Category</th>
<th>Year-Month</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2006-02</td>
<td>1.37</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2006-03</td>
<td>1.28</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2006-04</td>
<td>1.26</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2006-05</td>
<td>1.33</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2006-06</td>
<td>1.42</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2006-07</td>
<td>1.35</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2006-08</td>
<td>1.28</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2006-09</td>
<td>1.34</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2006-10</td>
<td>1.44</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2006-11</td>
<td>1.39</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2006-12</td>
<td>1.3</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2007-01</td>
<td>1.24</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2007-02</td>
<td>1.3</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2007-03</td>
<td>1.38</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2007-04</td>
<td>1.36</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2007-05</td>
<td>1.37</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2007-06</td>
<td>1.21</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2007-07</td>
<td>1.2</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2007-08</td>
<td>1.18</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2007-10</td>
<td>1.2</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2007-11</td>
<td>1.29</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2007-12</td>
<td>1.14</td>
</tr>
</tbody>
</table>

Figure 4: The number of sale transactions in each month in 2006-2007 in Zipcode 33175, excluding outliers, and the median of their ratios of the sale price to the fixed objective metric of the county valuation as of 1/1/2021; months with less than six transactions (September 2007) are excluded.

To facilitate human comprehension of said average (or median) ratios, we can normalize them to a specific period (month) as the base, e.g., the beginning month of said timespan, i.e., by computing the Factor as the median Ratio of any given month divided by the median Ratio of the base period. Thereby average (or median) prices can be expressed as the percentage increase (or decrease) since the base month, as in the following figures.
<table>
<thead>
<tr>
<th>Zipcode</th>
<th>Property Type</th>
<th>Month</th>
<th>Number of Closings</th>
<th>Factor</th>
<th>Percentage Increase since 02/2006</th>
</tr>
</thead>
<tbody>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2022-06</td>
<td>7</td>
<td>1.37</td>
<td>37% (since 2006-02)</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2022-05</td>
<td>24</td>
<td>1.35</td>
<td>35% (since 2006-02)</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2022-04</td>
<td>22</td>
<td>1.34</td>
<td>34% (since 2006-02)</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2022-03</td>
<td>25</td>
<td>1.27</td>
<td>27% (since 2006-02)</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2022-02</td>
<td>25</td>
<td>1.18</td>
<td>18% (since 2006-02)</td>
</tr>
<tr>
<td>33175</td>
<td>Houses</td>
<td>2022-01</td>
<td>17</td>
<td>1.29</td>
<td>29% (since 2006-02)</td>
</tr>
</tbody>
</table>

Figure 5: Normalization of the median ratios (the realized prices divided by the 2021 county valuation) to Month 2006-02, i.e., dividing by the median Ratio of 2006-02, whereby the last column shows the percentage increase since 2006-02.

Figure 6: Normalization of the median ratios (the realized prices divided by the 2021 county valuation) of January-June 2022 to February 2006, i.e., dividing by the median Ratio of 2006-02; the last column shows the percentage increase since 2006-02.
For better understanding by users, said factors can be presented as a graph, as in the following figure. Said chart informs how property values in the locale changed over time. The locale can be of any size as long as there are enough sale transactions therein to make a statistically significant analysis. The example in the following figure shows entire Southeast Florida as one locale and differentiates two property categories: condominium units and single-family homes.
3. **PSEUDO-CODE**

1. $MLS := \text{database of all multiple-listing service real estate transactions in SE Florida}$

2. $State\_Parcels := \text{database of county valuations of all properties in Florida as of a fixed date, e.g., 1/1/2021}$

3. $Allreal := \text{inner join on the field of FOLIO\_NUMBER of the MLS and Parcel databases}$:
   
   $MLS \ [\text{FOLIO\_NUMBER}] \ State\_Parcels;$
   
   and $projection$ of said join to all the fields on of $MLS$ plus the field $Just\_Value$ from $State\_Parcels$, i.e.:
   
   $Allreal := select \ MLS.*, \ State\_Parcels.Just\_Value from \ MLS, \ State\_Parcels where \ MLS.Folio\_nbr = State\_Parcels.Folio\_nbr$

4. $Zipcodes := \text{all the zip codes in Allreal, i.e.}$:
   
   $Zipcodes := select unique \ Zipcode \ from \ Allreal$

5. $for \ every \ \text{zipcode} \ in \ Zipcodes \ do \{}$

5.1. $Sub\_Allreal := select * from \ Allreal where \ Allreal.Zipcode = \text{zipcode}$

5.2. $Months := select unique (Closing\_Date \ as \ 9$
yyyy-mm-dd).substring(1,7) from Sub_Allreal

5.3. for each month in Months
    let Factor[zipcode,month] :=
    select median(Closing_Price/Just_Value)
    from Sub_Allreal
    where Closing_Date is within month

5.4. reference_month := minimum(Months) (Any month can be chosen to serve as the reference, in particular, it could be the minimum (earliest) month or the maximum (latest) month.)

5.5. Display or plot
    Factor[zipcode,*] / Factor[zipcode,reference_month]

4. ALTERNATIVE MODEL WITH CONTRACT-PENDING DATES

The closing date of property sale transactions has an imperfection in its utility to assess the contemporary market sentiment. That is because the market sentiment is manifested at the time of the execution of a contract for purchase and sale between the buyer and the seller, while the closing of the transaction typically occurs a month or a couple of months later. To capture the timeliness of the market sentiment more precisely, we can look at transactions that have closed, but we date them at the purchase contract’s effective date rather than at the closing date. Said purchase contract date can typically be obtained from MLS (multiple listing service) data sources (where it is often called the “Pending Date,” i.e., the date the property went under a purchase contract and became pending closing), like in the following figure.
By reanalyzing the same data for sales closed between January 2006 and June 2012, we get a chart more accurately showing the timely market sentiment during most periods, as in the following figure.

Figure 9: MLS data showing the Contract-Pending Date, in addition to the Closing Date, as well as the ratio of the closed sale price to the 2021 county valuation.
Although in this model we have more accurate market sentiment analysis in most periods, we do have noise bias at the edges. The two rightmost data points in this example aggregate properties closed by June 2022 but contracted for purchase in May or June 2022 (because the chosen timespan in this example is user-defined as properties closed from 1/2006 to 6/2022). Because the time elapsing between the contract date and the closing date in said May and June’s data is very short, these data points are biased towards cash sales (not contingent on mortgages), which often allow the buyer to negotiate lower prices. This bias can be excluded by disregarding the rightmost edge of the chart. There is also a bias noise at the left edge of the chart because the leftmost points include few but unusual transactions with contract dates as early as April 2005 that were closed in January 2006 or later. This bias can be excluded by disregarding the transactions where the purchase contract date is prior to the beginning of the user-chosen timespan (in this example, January 2006).

5. HIERARCHY OF LOCALES

Large locales, e.g., states and metropolitan areas, can be partitioned into smaller locales, e.g., townships and zipcodes, thus enabling the comparison of a locale to its neighbors as well as to its subsuming locales, as follows.
Figure 11: Partitioning Southeast Florida into a hierarchy of smaller locales
6. **STATISTICAL AGGREGATORS AND OUTLIERS**

A representative statistical aggregator function is a function that matches any set of numbers to a single number intended to be a typical representative of said set. Examples of representative statistical aggregator functions are:

- Median (“Pure Median”)
- Average (“Pure Average”)
- Average of the input set’s elements excluding the lowest 10% and the highest 10% of said set
- 0.5*Median+0.5*Average
- 0.3*Median+0.7*(Average of the input set’s elements excluding the lowest 5% and the highest 9% of said set)
- Average of the input set’s elements, excluding those elements that are outside predefined outlier thresholds of minimum 0.5 and maximum 1.5.

The present method involves the computation of a representative statistical aggregator function of all the purchase transactions in a given locale during a given period.

The easiest such aggregator function to compute is Pure Average. Among various statistical concerns with the Pure Average function, it may deliver significantly misleading results if the input data is not pre-cleansed off outliers. The Pure Median aggregator is more resilient to outliers, yet it still can benefit from the pre-cleansing of outliers. Outliers can be the result of

(a) erroneous data entry or
(b) the inclusion of esoteric transactions.

From the data cleansing algorithms’ point of view, there are several types of outlier cleansing that can be applied to a dataset of said ratios between transactional prices and the fixed-date objective valuation.

- Fixed threshold: disregard transactions with ratios outside of a given range, e.g., the range of 0.5 to 3.0.
- Percentage threshold: for a given category of properties, locale, and period, disregard certain percentages of the lowest and the highest ratios, e.g., the lowest 10% and the highest 5%.
- Statistically insignificant periods: for a given category of properties, locale, and period, if the number of the otherwise qualified transactions in the period is very small, e.g., less than 6, disregard all these transactions, i.e., skip this period for this locale (and for trend presentation purposes, interpolate this period form neighboring periods).

- Date-dependent threshold: for transaction dates far removed from the fixed year of the valuation, allow more liberal thresholds than those close to the valuation year. For example, if the objective valuation date is 1/1/2021, then for transactions in year y, where y<2021, e.g., y=2010, set the minimum threshold to 0.7-0.05*(2021-y).
- Semantic outliers that involve analysis of additional data fields, for example:
  (a) If there is a data field indicating that this is a foreclosure sale, disregard the transaction for being esoteric, with the expected price being too low.
  (b) Likewise, for short sales.
  (c) If there is a data field showing when the house was built (what in the governmental language is called “year of property improvement”), then disregard the transactions where said improvement date falls in between the transaction date and the fixed objective valuation date – this would prevent, e.g., the incorrect relating of the sale price of a building to the appraised value of bare land before the building was built).

7. **PROTOTYPE DEPLOYMENT**

We have deployed a system for Southeast Florida based on the algorithms presented here. Using county and MLS data, the system computes the value trend using transaction closing dates [5] and contract dates [6]. The model is computed for nested areas down to a zipcode and the category of condos vs. single-family homes. For example, the price trend of condominium apartments in Zipcode 33140 is at [7], and for houses is at [8]. The contract-date model for houses and condos in Zipcode 33140 is at [9] and [10].

**AUTHORS’ CONTRIBUTIONS**

Conceptualization: Rishe; Methodology: Rishe and Adjouadi; Investigation: Rishe, Tamir, and Adjouadi; Writing: Rishe and Tamir; Statistics: Tamir; Funding acquisition: Rishe and Adjouadi. All the authors of this paper concur with its content and consent to its publication.
ACKNOWLEDGMENT

This material is based in part upon work supported by the National Science Foundation under Grant Nos. MRI20 CNS-2018611, MRI CNS-1920182, B-BROIPS FDEP C-2104, and DHS E2055778.

REFERENCES


Risque [http://cake.fiu.edu/Rishe] is the inaugural outstanding university professor for Florida International University (FIU). He is a professor at FIU's Knight Foundation School of Computing and Information Sciences. Rishe is the Director of FIU's High-Performance Database Research Center. From 2013 to 2022, Rishe was the executive director of a multi-university center for Advanced Knowledge Enablement sponsored by the National Science Foundation's industry-university cooperation program. Rishe's research has been funded by the U.S. Government and Industry at over $50 million. Rishe's inventions include 26 U.S. patents. Rishe is a Fellow medalist of the National Academy of Inventors. In 2021, Rishe received the IBM Global University Programs award. Rishe's publications include 440 papers and six books. Rishe's pioneering geospatial system TerraFly has been highlighted by the National Science Foundation in its reports to U.S. Congress.

Tamir [https://userweb.cs.txstate.edu/~dt19/] is an associate professor in the Department of Computer Science at the Texas State University. He obtained the PhD-CS from Florida State University and the MS/BS-EE from Ben-Gurion University, Israel. He has managed applied research and design in DSP Core technology in Motorola-SPS/Freescale. He has served as an assistant/associate professor at the CS Department at Florida Tech, has served as a member of the MPEG committee and has worked in the Applied Research Division, Tadiran, Israel.

Adjouadi is the WARE Distinguished Professor at FIU's Department of Electrical and Computer Engineering. Since 1993, Adjouadi has directed the FIU Center for Advanced Technology and Education [http://cate.fiu.edu]. Adjouadi’s doctorate is from the University of Florida (1985).