The potential of big housing data: an application to the Italian real estate market*

Banca d’Italia, Big Data Workshop, 27/03/2018

Michele Loberto¹, Andrea Luciani¹, Marco Pangallo²

1 Banca d’Italia, Directorate General for Economics, Statistics and Research
2 Institute for New Economic Thinking at the Oxford Martin School and Mathematical Institute, University of Oxford

*The views expressed in this paper are those of the authors and do not necessarily reflect the views of Banca d'Italia
Why studying housing markets?

• Current economic situation: business activity of the construction industry, consumption through wealth effects (housing is the “democratic asset”)
  - Existing data available about 3 months later; limited hedonic-corrected indexes; no info on demand & supply

• Implications for financial stability
  - US housing bubble different across segments (Piazzesi et al., 2015, AER)

• Long-term effects in the labor market: misallocation of workers in the most productive areas (Hsieh and Moretti, 2017)
  - Absence of microdata
Why studying housing markets? (and current limits with Italian data)

• Current economic situation: business activity of the 
construction industry, consumption through wealth effects (housing is the “democratic asset”)
  – Existing data available about 3 months later; limited hedonic-correction; no info on demand & supply

• Implications for financial stability
  – US housing bubble different across segments (Piazzesi et al., 2015, AER): very limited information on segments

• Long-term effects in the labor market: misallocation of workers in the most productive areas (Hsieh and Moretti, 2017)
  – Tradeoff spatial coverage-representativeness
Our data

• Housing sales advertisements (ads) posted on the website immobiliare.it
• 1.1 million ads posted 2015Q1 to 2017Q2 in all province capital cities (30% of Italian population)
  • Here focus on “standard” dwellings for sale with full ownership
  • Ads can also refer to offices and shops, rents, foreclosure auctions, buildings under construction, partial ownership, usufruct, etc.
Our data

• Housing sales advertisements (ads) posted on the website immobiliare.it
• 1.1 million ads posted 2015Q1 to 2017Q2 in all province capital cities (30% of Italian population)
• Here focus on “standard” dwellings for sale with full ownership
• Ads can also refer to offices and shops, rents, foreclosure auctions, buildings under construction, partial ownership, usufruct, etc.
Our data

• For each ad we know:
  – Date posted, date removed
  – Geolocation
  – Floor area, rooms, bathrooms, floor, energy class, etc.
  – Garden, garage, elevator
  – Textual description & figure hash codes
  – (Asking) price, number of clicks every Friday of each week

• Spatial matching with 2011 census and tax agency (OMI) data
Trilocale via Etruria, Roma
120 m² | 3 locali | 1 bagno | € 465.000

Id: 57700898

Dati principali

Riferimento e Data annuncio  CDC2908 (1548035) - 02/09/2016
Contratto  Vendita
Tipologia  Appartamento
Superficie  120 m²
Locali  2 camere, 1 bagno, cucina abitabile
Piano  2° di 5 piani, con ascensore
Disponibilità  Libero
Tipo proprietà  Intera proprietà, classe immobile media

Caratteristiche

Parzialmente arredato  Cantina  Giardino comune  Balcone  Terrazzo

Descrizione

Le spese condominiali ammontano a c. a. €950 annue più €520 annue di riscaldamento
Il prezzo richiesto è di € 465.000,00.
Rif. CDC0308
Main issue: duplicates

• A seller can entrust more than one agency for the sale of the dwelling
• The same agency might delete an ad and create a new identical one for it to seem more recent
• Other possibilities
Is this really a problem? Yes

- Volume of transactions in 2016

<table>
<thead>
<tr>
<th>City</th>
<th>IMM</th>
<th>OMI</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turin</td>
<td>20263</td>
<td>12322</td>
<td>164.4</td>
</tr>
<tr>
<td>Genoa</td>
<td>10358</td>
<td>6601</td>
<td>156.9</td>
</tr>
<tr>
<td>Milan</td>
<td>40342</td>
<td>21909</td>
<td>184.1</td>
</tr>
<tr>
<td>Bologna</td>
<td>7655</td>
<td>5507</td>
<td>139.0</td>
</tr>
<tr>
<td>Florence</td>
<td>12833</td>
<td>4786</td>
<td>268.1</td>
</tr>
<tr>
<td>Rome</td>
<td>73070</td>
<td>30173</td>
<td>242.2</td>
</tr>
<tr>
<td>Naples</td>
<td>9764</td>
<td>6650</td>
<td>146.8</td>
</tr>
<tr>
<td>Palermo</td>
<td>5504</td>
<td>4718</td>
<td>116.7</td>
</tr>
</tbody>
</table>

Table 2: Transactions. Comparison between ads and OMI data.
**Pre-processing of the ads**

💰 € 196,000

70 m²

Maintenance: Excellent

**Pairwise identification of duplicates**

💰 € 196,000

70 m²

Maintenance: Excellent

\[ P(dup) = 0.62 \]

C5.0 classification tree

**Clustering pairs of duplicates and aggregation of information**

💰 € 194,000

60 m²

Maintenance: Excellent

(2.2; 5; -1.8; 9; -6)

doc2vec

(2.5; 7; -2.2; 8; -2)
Did it work?

- 1.1 million ads -> 650 thousand housing units
- Volume of transactions in 2016

<table>
<thead>
<tr>
<th>City</th>
<th>IMM</th>
<th>OMI</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turin</td>
<td>7615</td>
<td>12322</td>
<td>61.8</td>
</tr>
<tr>
<td>Genoa</td>
<td>4816</td>
<td>6601</td>
<td>73.0</td>
</tr>
<tr>
<td>Milan</td>
<td>12570</td>
<td>21909</td>
<td>57.4</td>
</tr>
<tr>
<td>Bologna</td>
<td>2866</td>
<td>5507</td>
<td>52.0</td>
</tr>
<tr>
<td>Florence</td>
<td>4122</td>
<td>4786</td>
<td>86.1</td>
</tr>
<tr>
<td>Rome</td>
<td>22103</td>
<td>30173</td>
<td>73.3</td>
</tr>
<tr>
<td>Naples</td>
<td>3832</td>
<td>6650</td>
<td>57.6</td>
</tr>
<tr>
<td>Palermo</td>
<td>2427</td>
<td>4718</td>
<td>51.4</td>
</tr>
</tbody>
</table>
Comparison and validation

Aggregation level: city

(e) Prices - Levels

ADS = -25.71 + 1.03 * Dwellings
$R^2 = 1$

(f) Prices - Growth rates

ADS = -0.05 + 0.91 * Dwellings
$R^2 = 0.65$
Comparison and validation

Aggregation level: OMI microzone

(e) Prices - Levels

(f) Prices - Growth rates

Same pattern for volume of transactions and stock of dwellings for sale
Comparison and validation

Validation OMI-Immobiliare.it

(a) Prices

OMI = 20.817 + 0.883 * IMM
R^2 = 0.83

(b) Transactions

IMM = -43.07 + 0.67 * OMI
R^2 = 0.97

Dwellings out of the market
Applications: nowcasting
Applications: nowcasting
### Applications: hedonic regression

#### Table 10: Hedonic regression

<table>
<thead>
<tr>
<th></th>
<th>Italy</th>
<th>Price per m² Rome</th>
<th>Milan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of bathrooms</td>
<td>154.577*** (3.743)</td>
<td>72.178*** (8.034)</td>
<td>208.084*** (12.031)</td>
</tr>
<tr>
<td>Floor of the apartment</td>
<td>32.363*** (0.926)</td>
<td>51.505*** (2.077)</td>
<td>48.876*** (2.338)</td>
</tr>
<tr>
<td>Floor area [m²]</td>
<td>-1.281*** (0.048)</td>
<td>-2.811*** (0.102)</td>
<td>-0.141 (0.174)</td>
</tr>
<tr>
<td>Number of rooms</td>
<td>-22.093*** (2.103)</td>
<td>-48.384*** (4.904)</td>
<td>-21.860*** (8.086)</td>
</tr>
<tr>
<td>Air conditioning</td>
<td>106.661*** (3.499)</td>
<td>126.300*** (7.545)</td>
<td>122.863*** (10.211)</td>
</tr>
<tr>
<td>Balcony</td>
<td>-30.398*** (3.693)</td>
<td>11.483 (7.912)</td>
<td>0.764 (10.935)</td>
</tr>
<tr>
<td>Elevator</td>
<td>168.574*** (3.969)</td>
<td>144.780*** (9.708)</td>
<td>223.752*** (12.014)</td>
</tr>
<tr>
<td>Energy class; Ref: EG; Level: AB</td>
<td>360.102*** (7.560)</td>
<td>267.614*** (18.907)</td>
<td>393.328*** (21.635)</td>
</tr>
<tr>
<td>Energy class; Ref: EG; Level: CD</td>
<td>114.052** (5.742)</td>
<td>92.140** (20.624)</td>
<td>135.590*** (16.620)</td>
</tr>
<tr>
<td>Garage; Ref: No; Level: Single</td>
<td>120.817*** (5.838)</td>
<td>127.548*** (11.158)</td>
<td>108.439*** (27.483)</td>
</tr>
<tr>
<td>Garage; Ref: No; Level: Double</td>
<td>205.992*** (4.330)</td>
<td>319.284*** (10.700)</td>
<td>149.507*** (12.491)</td>
</tr>
<tr>
<td>Heating type; Ref: No; Level: Centralized</td>
<td>47.737*** (10.439)</td>
<td>156.469*** (29.439)</td>
<td>225.680** (88.304)</td>
</tr>
<tr>
<td>Heating type; Ref: No; Level: Autonomous</td>
<td>135.250*** (9.878)</td>
<td>222.994*** (29.050)</td>
<td>281.075*** (88.417)</td>
</tr>
<tr>
<td>Kitchen type; Ref: Kitchenette; Level: Small eat-in kitchen</td>
<td>-49.458*** (5.350)</td>
<td>-23.742*** (11.736)</td>
<td>-89.975*** (14.268)</td>
</tr>
<tr>
<td>Kitchen type; Ref: Kitchenette; Level: Large eat-in kitchen</td>
<td>-77.409*** (4.477)</td>
<td>-86.146*** (10.131)</td>
<td>-57.177*** (12.522)</td>
</tr>
<tr>
<td>Status; Ref: To renovate; Level: Good</td>
<td>192.238*** (5.661)</td>
<td>164.210*** (11.707)</td>
<td>186.031*** (16.280)</td>
</tr>
<tr>
<td>Status; Ref: To renovate; Level: Excellent</td>
<td>480.591*** (5.874)</td>
<td>462.987*** (12.252)</td>
<td>535.513*** (16.606)</td>
</tr>
<tr>
<td>Status; Ref: To renovate; Level: New</td>
<td>610.984*** (8.808)</td>
<td>479.285*** (21.080)</td>
<td>584.823*** (25.053)</td>
</tr>
<tr>
<td>Terrace</td>
<td>143.589*** (3.761)</td>
<td>229.414*** (8.255)</td>
<td>217.474*** (12.834)</td>
</tr>
<tr>
<td>Utility room</td>
<td>-56.913*** (3.516)</td>
<td>-84.680*** (8.234)</td>
<td>-77.461*** (10.396)</td>
</tr>
<tr>
<td>Basement</td>
<td>11.039*** (3.660)</td>
<td>57.360*** (8.388)</td>
<td>24.695*** (9.814)</td>
</tr>
<tr>
<td>Porter</td>
<td>99.019*** (5.557)</td>
<td>73.887*** (10.685)</td>
<td>62.285*** (10.476)</td>
</tr>
<tr>
<td>Constant</td>
<td>718.810*** (241.357)</td>
<td>2,026.932*** (724.520)</td>
<td>457.625 (749.341)</td>
</tr>
</tbody>
</table>

**Note:**

*p<0.1; **p<0.05; ***p<0.01

OMI micro-zone and quarter dummies
Applications: segmentation

(a) Comparison
Applications: segmentation

(a) Comparison

(b) Floor area

(c) Energy class

(d) Maintenance status
Applications: clicks and demand tightness
Tightness as leading indicator of prices

\[
\log(P_{i,t}) = \alpha + \beta_1 \log(D_{i,t-1}) + \beta_2 \log(D_{i,t-2}) + \gamma T + \delta_0 \log(P_{i,t-1}) + \delta X_i + \epsilon_{i,t}
\]

- \( i \) is an OMI micro-zone (neighborhood)
- \( t \) is a quarter
- \( D \) is tightness (number of clicks per ad)
- \( T \) is quarter fixed effects
- \( X_i \) is OMI micro-zone characteristics (from census). No chance to use OMI fixed effects (so far).

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>Price per m2 (t) [log]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price per m2 (t-1) [log] (( \delta_0 ))</td>
<td>0.506*** (0.010)</td>
</tr>
<tr>
<td>Tightness (t-1) [log] (( \beta_1 ))</td>
<td>0.044*** (0.013)</td>
</tr>
<tr>
<td>Tightness (t-2) [log] (( \beta_2 ))</td>
<td>0.048*** (0.013)</td>
</tr>
<tr>
<td>Constant (( \alpha ))</td>
<td>3.211*** (0.120)</td>
</tr>
</tbody>
</table>

| Observations | 6,288 |
| R^2 | 0.811 |
| Adjusted R^2 | 0.807 |
Conclusion

• Online granular housing data have large potential both for policy and for research, for the analysis of the current economic situation and for long run economic outlook. They address many shortcomings of existing data sources (at least for Italy).

• **Next steps**: matching with notary deeds data to obtain transaction prices and with mortgage data to obtain loan-to-value

• At INET we build Agent-Based Models (ABMs). Using these datasets would allow ABMs to deliver their full potential in quantitative forecasting. (Geanakoplos, Axtell, Farmer, Howitt *et al.*, 2012, *AER P&P*; Haldane and Turrell, 2017, *Oxrep*)
Thank you for your attention!