Mining Heterogeneous Urban Data at Multiple Granularity Layers

Antonio Attanasio

Supervisor: Prof. Silvia Chiusano
Co-supervisor: Prof. Tania Cerquitelli
Urban data analytics

Added value urban services for citizens

- Data provision
- Analytics
- Citizens
- Services
- Data collection
- Feedback
- Service processing
- Exploitation

Credits: realityplusmag.com
Urban data analytics

Analysis of data collected in the urban context

<table>
<thead>
<tr>
<th>BENEFITS</th>
<th>ISSUES</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEFINE BETTER PLANNING POLICIES</td>
<td>VOLUME</td>
</tr>
<tr>
<td>OPTIMIZE RESOURCES UTILIZATION</td>
<td>HETEROGENEITY</td>
</tr>
<tr>
<td>INCREASE THE WELL-BEING OF CITIZENS</td>
<td>SPARSITY</td>
</tr>
<tr>
<td></td>
<td>INACCURACY</td>
</tr>
</tbody>
</table>

need for

Efficient data analytics algorithms

Innovative storage and computing infrastructures
PhD research activity

URBAN DATA ANALYTICS
Heterogeneous urban data management and analysis

BUILDINGS
ENERGY EFFICIENCY
Operational rating with real energy-related data
Real time predictive analytics
Asset rating modeling

PEOPLE'S PERCEPTION
Characterization of relevant topics from social networks
Urban data management and analysis

**SMART CITY CONTEXT**

*Manifold data from heterogeneous sources*

- Different spatio-temporal representations
- Different integration techniques

**MULTIPLE SPATIO-TEMPORAL PATTERNS**
Urban data management and analysis

Original contribution

- Distributed on-the-fly space-time data aggregation with MapReduce
- Distributed data analysis with MapReduce
- Fast exploration of different space and time resolutions
MuSTLE: Multiple Spatio-Temporal Layers Explorator

Data at highest resolution

Dynamic datasets

Time-space heterogeneity

Data volume

Distributed doc-oriented DB

Sharded Cluster

Need multiple time-space resolutions

Test different resolutions on the fly

MapReduce linear scalability

Elaboration time

ANALYSIS

AGGREGATION

STORAGE

DATA SOURCES

Dynamic datasets

Data volume

Distributed doc-oriented DB
Results of analysis

**DATA:** mobility, weather, air pollutants, urban facilities

### CORRELATION ANALYSIS

<table>
<thead>
<tr>
<th></th>
<th>Temperature</th>
<th>Humidity</th>
<th>Precipit.</th>
<th>Wind speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy cons.</td>
<td>-0.82</td>
<td>0.42</td>
<td>0.22</td>
<td>-0.59</td>
</tr>
<tr>
<td>CO</td>
<td>-0.56</td>
<td>0.14</td>
<td>-0.18</td>
<td>-0.56</td>
</tr>
<tr>
<td>NO₂</td>
<td>-0.40</td>
<td>-0.04</td>
<td>-0.23</td>
<td>-0.53</td>
</tr>
<tr>
<td>NO</td>
<td>-0.42</td>
<td>0.04</td>
<td>-0.22</td>
<td>-0.41</td>
</tr>
<tr>
<td>C₆H₆</td>
<td>-0.47</td>
<td>0.19</td>
<td>-0.23</td>
<td>-0.49</td>
</tr>
<tr>
<td>PM₁₀</td>
<td>-0.22</td>
<td>0.11</td>
<td>-0.31</td>
<td>-0.36</td>
</tr>
</tbody>
</table>

### REGRESSION ANALYSIS

- **CO [µg/m³]**
- **Energy per volume unit [Wh/m³]**

### KPIs COMPUTATION

(e.g., Total Heat Loss Coefficient)
MuSTLE computing performance

Computation of KPIs of buildings thermal *energy efficiency* almost linear speed-up

- > 2k buildings
- > 50M records
- 1 to 5 computing nodes

---

BUILDINGS ENERGY EFFICIENCY
Buildings energy efficiency

**DATA**
- Real energy measures
- Building features

**ANALYTICS**
- Descriptive
- Operational rating
- Asset rating modeling
- Predictive
- Real time predictive analytics
Buildings energy efficiency

DATA

Real energy measures

Building features

ANALYTICS

Descriptive

Operational rating

Predictive

Real time predictive analytics

Asset rating modeling
Operational rating with real energy data

Motivation

- Support the management of energy systems
- Help reducing inefficiency and energy waste
- Optimal sizing of the Heating Distribution Network
Context for operational rating

MONITORED DATA

Weather:
Outdoor temperature, humidity, precipitations, wind speed, etc.

Heating system:
Energy, power, water temperature, etc.

Environment:
Indoor temperature, humidity

Contextual data:
location of Buildings and sensors
Building volume, floor area, type, etc.

Credits: withenergy.co.uk
Operational rating with measured data

Original contribution

- KPIs at different space-time granularities computed with the MuSTLE framework

- Novel algorithm to predict instantaneous power demand values


A. Acquaviva; D. Apiletti; A. Attanasio; E. Baralis; L. Bottaccioli; S. Chiusano; E. Macii; E. Patti, Forecasting Heating Consumption in Buildings: a Scalable Full-Stack Distributed Engine. In: Expert Systems with Applications (SUBMITTED)
Real time predictive analytics

Prediction of buildings thermal power demand

- buildings have different daily \textit{heating cycles}
- during \textit{steady state} power is highly variable
- during \textit{transient state} peak power value is hardly predictable
Real time predictive analytics

Approach composed by three algorithms

- *Status and Outlier Detection (SOD)*: predicts the operational state
- *Peak Detection (PD)*: predicts the peak power value in the ON-line transient state
- *Power Prediction (PP)*: predicts the power profile in the ON-line phase
Power Prediction

- **Assumption:** power exchange correlated with surrounding weather conditions

- **Method:** building models based on linear regression (Energy Signature)
  - **Target variable:** power
  - **Explanatory variables:** temperature, humidity, wind speed, precipitations, pressure
  - One building model for each operational phase

- **Real time approach:** building models periodically updated with new measures
Power Prediction

- **Model learning**
  - Training window
  - Slot duration
  - Current time
  - Prediction target
  - Prediction horizon

- **Prediction at time** $t_j$
  - Predicted power value at $t_j$
  - Forecasted value of weather feature $i$ at $t_j$
  - Weight for weather feature $i$ computed during **model learning**

Array of weights
$W = [w_1, ..., w_n]$

For linear regression
$\hat{y} = \sum_{i=0}^{n} w_i \cdot x_i$

OUTPUT
Experimental results

Input data:
- 12 buildings with different heating cycles
- Original time granularity: 5 minutes

Results:
- Best time granularity: 30 minutes
- Best training window: 7 days

<table>
<thead>
<tr>
<th></th>
<th>Overall</th>
<th>1st cycle</th>
<th>2nd cycle</th>
<th>3rd cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SMAPE</strong></td>
<td>9.62%</td>
<td>10.80%</td>
<td>10.41%</td>
<td>9.51%</td>
</tr>
</tbody>
</table>

![Graph 1](image1.png)

![Graph 2](image2.png)
Asset rating with building features

Motivation

- **Building designing**: how building features affect the energy efficiency

- **Existing buildings**: estimate the improvements of a refurbishment plan
Buildings Energy demand modeling

Objective of the activity

- Quantify and explain building efficiency based on data from *Energy Performance Certificates*
  - Building geometric features
  - Physical features of building envelope
  - Building historical info
  - Energy related variables
- **Estimation** of Primary Energy Demand for space heating (*PED*ₕ)
- **Description** of its relationship with building features
Buildings Energy demand modeling

Original contribution

- **Methodology:** two-layer approach to characterize the energy demand of buildings
- **Analysis:** interpretable classification model

Buildings Energy demand modeling

Two-layers approach

- Segment estimation (classification)
- Local energy demand prediction (regression)

Algorithms

- Artificial Neural Networks
- Support Vector Machines
- Reduced Error Pruning Tree
- Random Forest
Buildings Energy demand modeling

Segment estimation: classification results

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>ANN</th>
<th>REPT</th>
<th>RF</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy [%]</td>
<td>67.51</td>
<td>82.03</td>
<td>85.67</td>
<td>67.24</td>
</tr>
</tbody>
</table>

- **OPAQUE TRANSMISSIONS** ($U_o$)
  - $< 0.56$
  - $\geq 0.56$
    - **CONSTRUCTION YEAR** ($y_c$)
      - $< 2007$
        - **ASPECT RATIO** ($R$)
          - $< 0.63$
            - AVG EFFICIENCY SPACE HEATING ($\eta_p$)
              - $< 0.82$
              - $\geq 0.82$
          - $\geq 0.63$
        - **OPAQUE TRANSMISSIONS** ($U_o$)
      - $\geq 2007$
    - **ASPECT RATIO** ($R$)
      - $< 0.50$
        - AVG EFFICIENCY SPACE HEATING ($\eta_p$)
          - $< 0.74$
          - $\geq 0.74$
      - $\geq 0.50$
        - AVG EFFICIENCY SPACE HEATING ($\eta_p$)
          - $< 0.77$
          - $\geq 0.77$

- **TRANSPARENT TRANSMISSION** ($U_w$)
  - $< 2.12$
    - **ASPECT RATIO** ($R$)
      - $< 0.55$
        - AVG EFFICIENCY SPACE HEATING ($\eta_p$)
          - $< 0.82$
          - $\geq 0.82$
      - $\geq 0.55$
    - **TRANSPARENT TRANSMISSION** ($U_w$)
      - $< 2.15$
        - **ASPECT RATIO** ($R$)
          - $< 0.74$
            - AVG EFFICIENCY SPACE HEATING ($\eta_p$)
              - $< 0.74$
              - $\geq 0.74$
          - $\geq 0.74$
        - **TRANSPARENT TRANSMISSION** ($U_w$)
          - $< 2.15$
            - **ASPECT RATIO** ($R$)
              - $< 0.74$
              - $\geq 0.74$
Buildings Energy demand modeling

Local energy demand prediction: regression results

<table>
<thead>
<tr>
<th></th>
<th>ANN</th>
<th>REPT</th>
<th>RF</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAPE [%]</td>
<td>27.02</td>
<td>16.64</td>
<td>16.89</td>
<td>21.52</td>
</tr>
</tbody>
</table>

- w.r.t. **single layer approach**
  - MAPE 16.64% vs 29.82% (REPT over the entire dataset)

- w.r.t. **related work**
  - Smaller set of explanatory variables
  - Algorithms trained in few seconds
  - Interpretable models provided to domain experts
PEOPLE'S PERCEPTION
Analysis of relevant topics from social networks

Motivation

- Understand collective dynamics of people’s interests and needs
  - Highly debated topics on social networks

- Devise targeted actions in the management of a city

- Characterize patterns of user interests across different cities
Analysis of relevant topics from social networks

SOCIAL NETWORK SITES

User-generated media

- text
- videos
- photos
- URLs

- time
- place
Analysis of relevant topics from social networks

Objective of the activity

- Characterization of popular topics from Twitter posts
- Discovery of text-spatio-temporal patterns of tweets
  - Cohesive clusters of tweets
  - Significant association rules
Analysis of relevant topics from social networks

Original contribution

- Extensible framework
  - Different clustering algorithms
  - Different distance measures

- TASTE distance measure: text, space, and time features

- Characterization through association rules

Xin Xiao, Antonio Attanasio, Silvia Chiusano, Tania Cerquitelli, Twitter data laid almost bare: An insightful exploratory analyser, In Expert Systems with Applications, Volume 90, 2017
Analysis of relevant topics from social networks

**Tweet representation**

<table>
<thead>
<tr>
<th>TEXT</th>
<th>Australia vs Netherlands I predict 3-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOCATION</td>
<td>52.051, -0.803</td>
</tr>
<tr>
<td>TIME</td>
<td>Wed Jun 18 15:55:55</td>
</tr>
</tbody>
</table>

**Twitter Streaming API**

**Cleaning (stopwords,...)**

**TF-IDF**

**Collection**

**Preprocessing**

**Tweets data set**

**Clustering analysis**

**Association rules analysis**

- FP-GROWTH algorithm
- Rules categorization

**Cluster set**

**Categorized association rules**
Analysis of relevant topics from social networks

**TASTE:** Text And Spatio-TEmporal distance

\[ d_{TASTE}(\tau_i, \tau_j) = d_W \cdot (k_s \cdot e^{ps \cdot d_s} + k_t \cdot e^{pt \cdot d_t}) \]
Analysis of relevant topics from social networks

Comparative analysis of distance measures

\[ d_{TASTE}(\tau_i, \tau_j) = d_W \cdot (k_s \cdot e^{p_s \cdot d_s} + k_t \cdot e^{p_t \cdot d_t}) \]

\[ d_{Kim}(\tau_i, \tau_j) = d_s \]
\[ d_{Arc}(\tau_i, \tau_j) = [\max(d_s, d_t)]^\beta \]
\[ d_{Lee}(\tau_i, \tau_j) = d_W \cdot e^{\frac{\zeta \cdot d_t}{M}} \]
\[ d_{Cun}(\tau_i, \tau_j) = w_W \cdot d_W + w_t \cdot d_t + w_s \cdot d_s + w_{so} \cdot d_{so} \]

\( \tau_i, \tau_j \): tweets representations
\( d_s \): Space distance
\( d_t \): Time distance
\( d_W \): Text distance
Analysis of relevant topics from social networks

Comparative analysis of distance measures

<table>
<thead>
<tr>
<th>Distance Measure</th>
<th>Mean time distance (min)</th>
<th>Mean space distance (km)</th>
<th>Mean text distance (rad)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TASTE</td>
<td>35</td>
<td>158</td>
<td>0.95</td>
</tr>
<tr>
<td>Kim</td>
<td>3,905</td>
<td>14</td>
<td>1.28</td>
</tr>
<tr>
<td>Arc</td>
<td>33</td>
<td>66</td>
<td>1.26</td>
</tr>
<tr>
<td>Lee</td>
<td>35</td>
<td>246</td>
<td>1.03</td>
</tr>
<tr>
<td>Cun</td>
<td>126</td>
<td>245</td>
<td>0.95</td>
</tr>
</tbody>
</table>

![Box plots showing the distribution of mean time, space, and text distance for different methods.]
Analysis of relevant topics from social networks

Clusters characterization with association rules

Rules with highest support and lift

- Different topics in different cities
- Different evolutions of topics over time in different cities

<table>
<thead>
<tr>
<th>RULE</th>
<th>CENTROID TIME</th>
<th>CENTROID LOCATION</th>
<th>SUPPORT</th>
<th>LIFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>{uruguay} ➔ {england}</td>
<td>2014-06-19</td>
<td>Perth (UK)</td>
<td>5%</td>
<td>2.4</td>
</tr>
<tr>
<td>{suarez,someone} ➔ {bite}</td>
<td>2014-06-25</td>
<td>Rugeley (UK)</td>
<td>3%</td>
<td>26.9</td>
</tr>
<tr>
<td>{nigeria} ➔ {argentina}</td>
<td>2014-06-25</td>
<td>Banning (US)</td>
<td>2.1%</td>
<td>7.2</td>
</tr>
</tbody>
</table>
Conclusion

Urban data management and analytics (MuSTLE)

- distributed data aggregation and analysis
- on the fly exploration of multiple combinations of space-time data granularities
- almost linear computational scalability

Credits: realtyplusmag.com
Conclusion

Urban services: buildings energy efficiency

- analysis of real consumption and of building features
- prediction of power levels and energy demand
- Actionable knowledge from interpretable models of energy demand

Credits: withenergy.co.uk
Conclusion

People's feedback from social media

- Extraction of clusters of tweets focused on relevant topics
- Space-time-text characterization of clusters
- New distance measure based on space, time and text features
Publications

CONFERENCES/WORKSHOPS


JOURNALS:
