Open City Data Pipeline - Collecting, Integrating, and Predicting Open Statistical City Data

Stefan Bischof\textsuperscript{1,2}, Christoph Martin\textsuperscript{1}, Axel Polleres\textsuperscript{1}, and Patrik Schneider\textsuperscript{1}

(1) Vienna University of Economics and Business, Austria
(2) Siemens AG, Austria

May 31, 2015
Overview

Introduction

Data Sources

Prediction of Missing Values

Results and Future Work
Introduction

Data Sources

Prediction of Missing Values

Results and Future Work
Motivation

Figure: European Green City Index 2009 (Siemens AG)
Aim

- A framework for **Smart City** applications (e.g., Green City Index)
- Gathering **performance indicators (KPIs)** for cities and published them as Open Data → timely
- Build on **open standards**, e.g. RDF, OWL, SPARQL
- “Semi-automatically” collect various Open Data Sources:
  - DBpedia
  - Eurostat – Urban Audit
  - United Nations Statistics Division
  - U.S. Census
- **Ontology-based** integration of these data sources.
Challenges

- Open Data sources for cities are around, but... the quality varies
- ... many (fragmented) access points, e.g., by cities, countries, NGOs
- ... are very heterogeneous:
  - Indicator specifications (time, units, etc.)
  - Format (CSV, RDF, etc.)
  - Licence (CC-BY, etc.)
  - Access points (plain download, SPARQL endpoint, etc.)

- Many gaps in the data sets → missing values
Open City Data Pipeline

- **System Architecture:**
  - Crawler
  - Wrapper
  - Ontology
  - Data Storage
  - Predictions, UI & LOD

- Storage is for now in Jena TDB
City Data Model

- Designed as an extensible City Data Ontology in $\mathcal{ALH}(D)$
- CityDataContext links: Spatial context (i.e. the city), Temporal context, Indicators, and data sources
- Data source specific indicators are modelled as sub-property to an abstract “super” city data indicators
Introduction

Data Sources

Prediction of Missing Values

Results and Future Work
Urban Audit

- Urban Audit collection is an initiative to assess the quality of life in European cities
- By the national statistical institutes and Eurostat
- Wide range of topics (e.g., on demography, environment, health, economics, and tourism) with 215 indicators

<table>
<thead>
<tr>
<th>Year(s)</th>
<th>Cities</th>
<th>Indicators</th>
<th>Filled</th>
<th>Missing</th>
<th>% of Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>177</td>
<td>121</td>
<td>2480</td>
<td>18937</td>
<td>88.4</td>
</tr>
<tr>
<td>2000</td>
<td>477</td>
<td>156</td>
<td>10347</td>
<td>64065</td>
<td>85.0</td>
</tr>
<tr>
<td>2005</td>
<td>651</td>
<td>167</td>
<td>23494</td>
<td>85223</td>
<td>78.4</td>
</tr>
<tr>
<td>2010</td>
<td>905</td>
<td>202</td>
<td>90490</td>
<td>92320</td>
<td>50.5</td>
</tr>
<tr>
<td>2004 - 2012</td>
<td>943</td>
<td>215</td>
<td>531146</td>
<td>1293559</td>
<td>70.9</td>
</tr>
<tr>
<td>All (1990 - 2012)</td>
<td>943</td>
<td>215</td>
<td>638934</td>
<td>4024201</td>
<td>86.3</td>
</tr>
</tbody>
</table>
Collected by the United Nations Demographic and Social Statistic Division ( UNSD)  
Focus of 650 indicators mainly on demographic and housing data  
Covers the entire world

<table>
<thead>
<tr>
<th>Year(s)</th>
<th>Cities</th>
<th>Indicators</th>
<th>Filled</th>
<th>Missing</th>
<th>% of Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>7</td>
<td>3</td>
<td>10</td>
<td>11</td>
<td>52.4</td>
</tr>
<tr>
<td>2000</td>
<td>1 391</td>
<td>147</td>
<td>7 492</td>
<td>196 985</td>
<td>96.3</td>
</tr>
<tr>
<td>2005</td>
<td>1 048</td>
<td>142</td>
<td>3 654</td>
<td>145 162</td>
<td>97.5</td>
</tr>
<tr>
<td>2010</td>
<td>2 008</td>
<td>151</td>
<td>10 681</td>
<td>292 527</td>
<td>96.5</td>
</tr>
<tr>
<td>2004 - 2012</td>
<td>2 733</td>
<td>154</td>
<td>44 944</td>
<td>3 322 112</td>
<td>98.7</td>
</tr>
<tr>
<td>All (1990 - 2012)</td>
<td>4 319</td>
<td>154</td>
<td>69 772</td>
<td>14 563 000</td>
<td>99.5</td>
</tr>
</tbody>
</table>
Combining different Data Sets

- Each data set has already a high ratio of missing values!
- Merging data sets with different indicators/cities adds sparsity:
  - Disjoint cities and aligning cities fails
  - Indicators by default are disjoint

<table>
<thead>
<tr>
<th>Data from Source 1</th>
<th>Data from Source 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vienna</td>
<td>Augsburg</td>
</tr>
<tr>
<td>Cars</td>
<td>655806</td>
</tr>
<tr>
<td>Nationals</td>
<td>1342704</td>
</tr>
<tr>
<td>Women per 1000 Men</td>
<td>109.8</td>
</tr>
</tbody>
</table>

Combined data from Source 1 and Source 2
Introduction

Data Sources

Prediction of Missing Values

Results and Future Work
Base Methods

- Our assumption: every indicator has its own distribution and relationship to others

- Basket of “standard” regression methods:
  - K-Nearest Neighbour Regression (KNN)
  - Multiple Linear Regression (MLR)
  - Random Forest Decision Trees (RFD)

- Normalized root mean squared error in % (RMSE%)
- Validation: Stratified tenfold cross-validation
A1: Building Complete Subsets

- Input matrix: rows are City/Year combinations, columns are Indicators
- Predictors directly taken from indicators
- Find best predictors by corr. matrix between target and other indicators
- Apply KNN, MLR, RFD methods and choose the best by RMSE%
- But need a complete matrix → cities with miss. values are deleted
A2: Principal Component Regression

- Instead of using indicators we use Principal Component (PC) Analysis
- Fill in missing data points with “neutral values” for the PCA
- Find best predictors by corr. matrix between target and PCs
- Apply KNN, MLR, RFD methods and choose the best one
- We now are able to fill in all missing values
Introduction

Data Sources

Prediction of Missing Values

Results and Future Work
Discussion of Results

Approach 1:
- Good result with a RMSE% of 0.25%
- But for many cities no predictions at all

Approach 2:
- Prediction of all missing values
- With a RMSE% of 3.29% result is not as good as Approach 1
- Still room for improvements
- Future use for the pipeline

Calculation using R (12 hours with an Intel Core i7 2.66GHz and 8 GB of RAM)
Publishing as Linked Data with a threshold RMSE% of 10% (28 indicators dropped)

Ask SPARQL queries

```
PREFIX rdfs: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX : <http://citydata.wu.ac.at/ns#>

SELECT DISTINCT ?city ?year ?indicator ?value WHERE { 
  GRAPH ?g { 
    ?spatialContext ?city ;
    ?property ?value ;
    ?date ?value ;
    ?indicator ?date ;
    BIND(?year) as ?year } 
  VALUES ?city { <http://dbpedia.org/resource/Vienna> <http://dbpedia.org/resource/Munich> } 
  VALUES ?indicator { ?population ;
    c:population_female } 
  ORDER BY ?city DESC(?year) ?indicator
}
```

Output: Text

If XML output, add XSLT style sheet (blank for none): 
□ Force the accept header to text/plain regardless. 
Get Results

Aachen

Go directly to one of the categories
- Culture and Recreation
- Demography
- Economic Aspects
- Environment
- Geography
- Social Aspects
- Training and Education
- Travel and Transport
- General

Culture and Recreation

Available beds per 1000

- 2004: 20.546 no (from http://citydata.wu.ac.at/ns#Prediction, predicted by k-nearest neighbor with an estimated error of 2.4173865752 %RMSE)
- 2005: 17.823 no (from http://citydata.wu.ac.at/ns#Prediction, predicted by k-nearest neighbor with an estimated error of 2.4173865752 %RMSE)
- 2006: 19.146 no (from http://citydata.wu.ac.at/ns#Prediction, predicted by k-nearest neighbor with an estimated error of 2.4173865752 %RMSE)
- 2008: 18.0 no (from http://epp.eurostat.ec.europa.eu/)
- 2009: 17.5 no (from http://epp.eurostat.ec.europa.eu/)
- 2010: 17.4 no (from http://epp.eurostat.ec.europa.eu/)
- 2011: 17.5 no (from http://epp.eurostat.ec.europa.eu/)
- 2012: 18.4 no (from http://epp.eurostat.ec.europa.eu/)

Cinema Attendance

- 2004: 1438509.364 persons (from http://citydata.wu.ac.at/ns#Prediction, predicted by multiple linear regression with an estimated error of 3.7662020028 %RMSE)
Ongoing / Outlook

- **Focus:** We developed a framework for filling in missing values for (statistical) city open data

- **Ongoing:**
  - Add new data sets, e.g., U.S. Census and individual cites (*QuerioCity*)
  - Add new methods, e.g., SVM and robust linear regression
  - Cross data set prediction
  - Learning Ontology Mappings from Indicator Values (*Explain-a-LOD*)

- **Future work:**
  - Introducing Time Series Analysis, maybe in combination with the existing methods
  - Mapping our Linked Data interface to standard vocabularies (PROV, RDF Data Cube)
  - Automatic validation and correction of predictions (value ranges)
  - Speed up the process by remembering the best settings for predictions
  - Integrating of OSM and Linked Geo Data