Where Do People Go When It Rains?

Marko Niinimaki
Helsinki Institute of Physics
CH-1211 Geneva 23,
Switzerland
man@cern.ch

Tapio Niemi
Helsinki Institute of Physics
CH-1211 Geneva 23,
Switzerland
tapio.niemi@cern.ch

Peter Thanisch
School of Information
Sciences
FIN-33014 University of
Tampere, Finland
pt@cs.uta.fi

ABSTRACT

The goal of this paper is two-fold. On one hand, we develop methods for integrating external data with the Mobile Data Challenge (MDC) data. We then analyse the resulting data using OLAP (On-Line Analytical Processing) and statistical tools. We describe the methods using which this has been achieved. On the other hand, after the integration we aim at answering the question: are there differences in peoples' mobility depending on the weather?

1. INTRODUCTION

The location data of mobile phones are increasingly being used in the analysis of mobility of their users [10, 3]. Noulas et al [9] have used the location-aware social networking services FourSquare to study the users' transition from one place to another; Miluzzo et al [5] have studied sharing the users' precense data though other social networks. Only a few studies [11], however, discuss a methodology of integrating mobile phone data with other data sources. This is our goal in the paper at hand, and we use weather data in addition to location data to demonstrate the method.

The data integration method is based on our earlier work of Semantic Web technologies in data integration [7, 8]. This makes it possible to utilise different external data sources that are in the RDF (Resource Description Framework [4, 2]) format, or can easily be converted to RDF. There are already lots of data available in RDF format and it is likely that this resource is growing rapidly. The main benefit in using Semantic Web technologies is that they support data integration based on the meaning of data. Thus, to use the data, the user does not need to know its structure, nor always even its location.

Our method presented in this paper will have the following steps:

1. Converting data sets into the RDF format if they are originally stored in some other format.
2. Integrating the data sets using a RDF query language.
3. Uploading the relevant subset of data into an OLAP engine or a statistical software for detailed analysis.
4. Analysing the data and testing hypotheses.

The method is tested by using detailed weather data from MeteoSwiss together with the MDC data. We use the integrated data to check how the current weather affects people's route choices or their traveling speed. The analysis is done using OLAP (On-Line Analytic Processing) methods [1]. Briefly, OLAP means presenting data in a form of a multi-dimensional cube (an OLAP cube) with hierarchical dimensions or coordinates. These coordinates give a structure to low-level data recorded in fact rows. OLAP allows us to analyse the data on different levels of details with different sets of dimensions. For example, we can check whether people move longer distances in warm weather and when 'drill-down' to see if there are some differences in the distance between different months.

The rest of the paper is organized as follows. In Section 2, we discuss the methodology of data integration using RDF. Section 3 shows how the integrated data can be analysed using the OLAP method. Finally, Section 4 contains conclusions.

2. DATA INTEGRATION

2.1 MDC Data and Its Pre-processing

We got our data set from The Mobile Data Challenge (MDC data) [1] and MeteoSwiss (Meteo data) [2].

The MDC data contains data from 38 persons between 2009-10-08 - 2011-03-23. Not all dates within this time span are covered. The minimum span for a person is 71 days and the maximum 383 days. Since the GPS data was collected from different start and end dates, we have concentrated on the following time span that is long enough and shared by a reasonable number (17) of subjects: 12th Nov 2009 - 1st May 2010 (171 days). Not everyone had their GPS or phone on every day. A sample of a route during one day by one person is shown in Figure 1. It should be noted that the sample is quite small in comparison with other studies [10,

This material was prepared for the Mobile Data Challenge 2012 (by Nokia) Workshop; June 18-19, 2012; Newcastle, UK. The copyright belongs to the authors of this paper.

[1] http://research.nokia.com/page/12000
[2] http://meteoswiss.ch

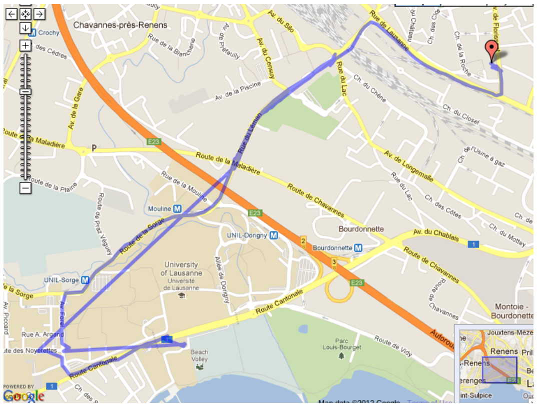

Figure 1: An example route

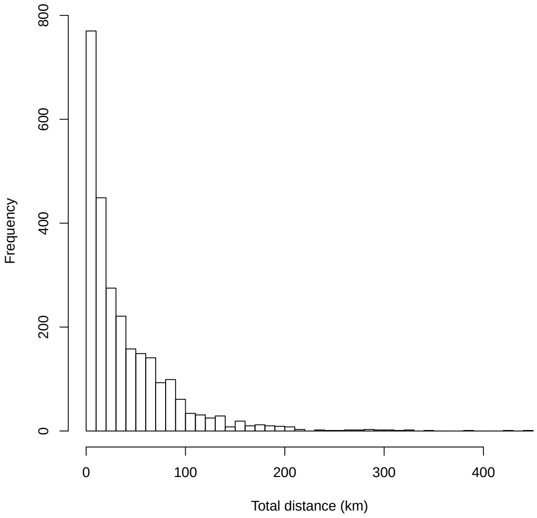

Figure 2: Daily distances

3]. The fact that the data does not cover a full year makes seasonal comparisons "informal" at best.

We calculated some new variables based on the GPS coordinates. Altogether, there are ca. 167 000 recorded GPS events which we mapped to 13 areas based on postal codes. The postal codes were acquired by using a reverse geocoding tool at www.findlatitudeandlongitude.com.

We also calculated Distance-to-next GPS event for all events. Based on this we were able to calculate distances. For example, the smallest total distance covered by a person during data collection per day was 8 meters, the longest 445.6 km, and the median 22.9 km. A histogram is shown in Figure 2 (N=2636).

To find out how far each subject travelled from home, we checked if for a given subject, the last recorded GPS location of the day had invariably (every day) the same postal code. By this method, we roughly evaluated the home location for each subject.

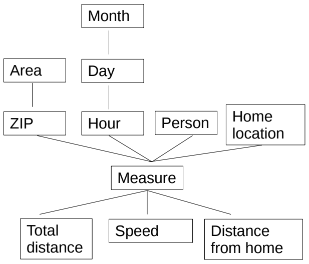

Figure 3: Mobility data in OLAP form

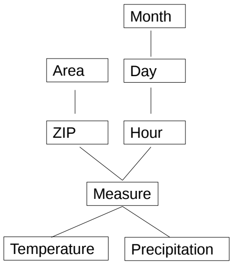

Figure 4: Weather data in OLAP form

2.2 Meteo Data

The weather data contains historical weather measurements such as temperature, precipitation, humidity, sun shine, wind, and air pressure. The data are recorded in each weather station in Switzerland during the same period as the MDC data. The frequency of data collection is hourly for some stations and once every 10 minutes for others. There are around 50 stations in Switzerland, so it is quite possible to get the local weather based on the recorded location of the mobile phone.

For data integration, we selected precipitation and temperature from the closest weather station according to recorded GPS locations in the MDC data.

2.3 Using RDF for Data Integration

The mobility data is converted to conform our RDF OLAP model [7] having the following dimensions: Person, Time, Area and measures Speed, and DistanceHome (Figure 3). The weather data has Location and Time as dimensions and Precipitation and Temperature as measures (Figure 4).

Both of these data sets are stored according to ontologies RDF XML files. The mobile data is stored in FactRow elements.

```
<FactRow rdf:about="FR-10-6.56495813539">
<rdfs:label>FR-10-6.56495813539</rdfs:label>
  <hasDimensionMember rdf:resource="#p10"/>
  <hasDimensionMember rdf:resource="#z9.4"/>
  <hasDimensionMember rdf:resource="#d1258038249"/>
  <hasDimensionMember rdf:resource="#t0"/>
  <hasDimensionMember rdf:resource="#t46.5255108"/>
  <hasMeasureMember rdf:resource="#MD10.1267237180638"/>
  <hasMeasureMember rdf:resource="#MLA469.5"/>
  <hasMeasureMember rdf:resource="#MLO46.525309987"/>
</FactRow>
```

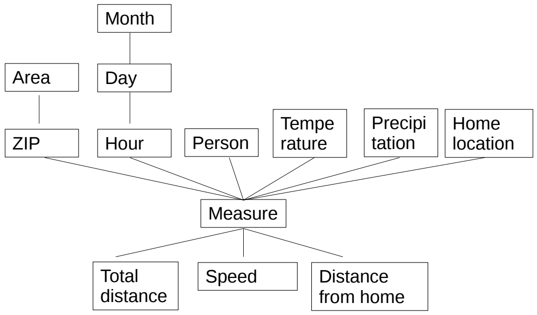

Figure 5: Integrated data in OLAP form

The weather data in the RDF format is as follows

```
 <weather:observation
  rdf:about="http://www.cern.ch/hiptek/
             weather.rdf#BER.200910080000">
     <weather:location>BER</weather:location>
     <weather:time>200910080000</weather:time>
     <weather:temperature>16.3</weather:temperature>
     <weather:rain>0.0</weather:rain>
 </weather:observation>
```

To populate the model in Figure 5 we integrate these two data sources using an RDF query:

```
SELECT Wtem, Wrain, F
FROM {W} weather:location {Wloc}
FROM {W} weather:time {Wtime}
FROM {W} weather:temperature {Wtem}
FROM {W} weather:rain {Wrain}
FROM {F} olapcore3:hasDimensionMember {D}
FROM {D} olapcore3:BelongsTo {X}
FROM {X} olapcore3:hasDimension {<file://#time>}
FROM {F} olapcore3:hasDimensionMember {A}
FROM {A} olapcore3:BelongsTo {Y}
FROM {Y} olapcore3:hasDimension {<file://#area>}
FROM {A} olapcore3:RollsUp {C}
WHERE Wloc=C AND Wtime=D
```

The query returns the local temperature and precipitation for each fact row in the database.

2.4 Analysis Software

We use the statistical analysis system R for our data analysis. R is not originally designed for OLAP analysis yet it has very suitable features for our purposes. We apply a so-called dimensionless OLAP model in which the whole data base is stored as one flat relation. This model is very flexible for ad-hoc type data analysis.

The reshape package of R supports OLAP analysis in an easy way. The cast function of the package produces an aggregated OLAP cube, that is, it aggregates and pivots the data and finally displays in a two-dimension table. It is also possible to define a sub set of data to be used in the analysis. This corresponds with the slice operation of OLAP. It is also possible to reuse the result of the cast operation as an input to future queries. This is done by the melt operation that re-formats the data back to its original form.

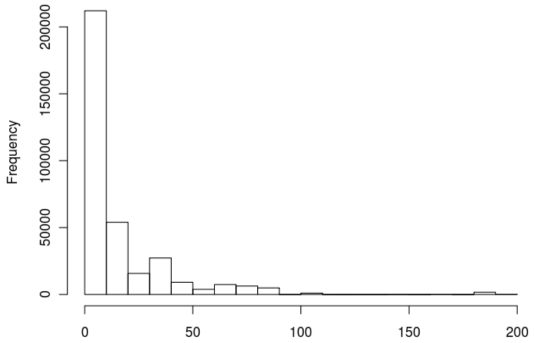

Figure 6: Distance from home (home location estimated from the day's first and last GPS location)

For example, compute daily average distances of all persons for Mondays and Tuesdays:

```
cast (daily, gps_weekday ~ ., fun.aggregate=mean,
      value=’dailydistance’,
      subset=gps_weekday == ’Monday’ |
             gps_weekday == ’Tuesday’)

  gps_weekday    (all)
1      Monday 33.61386
2     Tuesday 33.32333
```

Before the analysis, the RDF data must be uploaded to the analysis system. This is simply done by storing the data after integration in CSV format and when applying standard input commands of R.

3. DATA ANALYSIS

Before analysing the effect of precipitation, we check some basic temperature related indicators. The minimum temperature in places where persons have been during the sample period was -11 degree, maximum 34 degrees, mean 9.8 degrees. We note that temperature does not seem to affect distance from home (correlation of temperature and distance from home 0.008), total distance travelled during the day (correlation 0.058), or speed (correlation 0.024).

We start the mobility analysis by basic statistics. The mean of the distance covered by a person in one day is 39.7 km (max 445.6), mean of distance between 2 consecutive observations 0.16710, and mean of distance from home 15.07 (histogram in Figure 6). To analyse if persons move different distances in different weekdays is done by using the following query:

```
cast (daily, . ~ gps_weekday,
      fun.aggregate=mean, value=’dailydistance’)

  value   Monday   Tuesday  Wednesday   Thursday
1 (all) 33.61386  33.32333    35.5133   38.85541

 Friday   Saturday   Sunday
40.89312  48.43001  51.18817
```

The results indicate that there is no significant difference

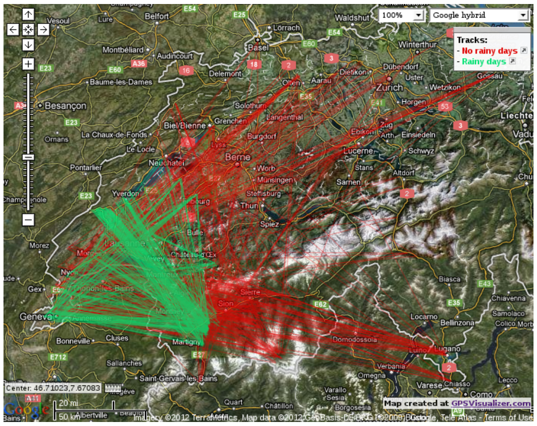
Tracks:
- No rainy days
- Rainy days
Map created at GPSVisualizer.com
Imagery ©2012 TerraMetrics, Map data ©2012 GeoBasis-DE/BKG (©2009), Google, Tele Atlas

Figure 7: People seem to travel more when the weather is good

by weekdays in the distance covered by persons but during weekends the distance gets longer.

Within the period of 171 days, there were hardly any days that were completely rainless everywhere where the subjects went (actually 1.1.2010 was rainless according to our data, but only one person's phone was active during that day). Similarly, every day during the sample period at least one person enjoyed a rainless day.

The "rainy day" variable is an aggregation from the 10 minute precipitation measurement: if there is any precipitation at any time during the day, "rainy day" is TRUE. Altogether, in our sample there were 132 000 GPS measurements during rainy days and 494 000 during clear days.

Precipitation in general seems to make people move less, as shown in Figure 7. This figure was composed by taking one point at intervals of 10 points of observations from each of the subjects. However, when computing daily average distance for rainy and clear days, the average distance moved in rainy days is longer. T-test gives a p-value 0.001107 but comparing histograms in Figures 8 and 9, we do not notice a clear difference in distributions.

```
cast (daily, . ~ rainyday,
      fun.aggregate=mean, value=’dailydistance’)

  value    FALSE     TRUE
1 (all) 38.24011 46.71316


cast (daily, . ~ rainyday,
      fun.aggregate=median, value=’dailydistance’)

  value    FALSE     TRUE
1 (all) 21.88617 30.25524
```

But is there really a correlation between daily precipitation and the daily distance travelled? In fact, the correlation between rain and distance is positive, 0.12.

Let us see what is really happening. On weekdays, people travel a bit more if it is rainy/snowy. However, rainy or snowy Saturdays make people move!

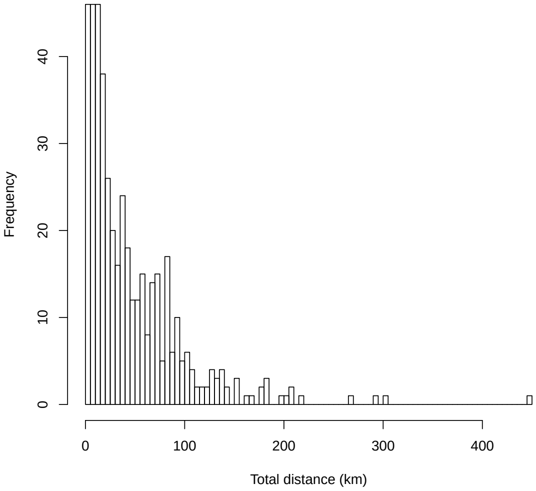

Figure 8: Daily distances on rainy days

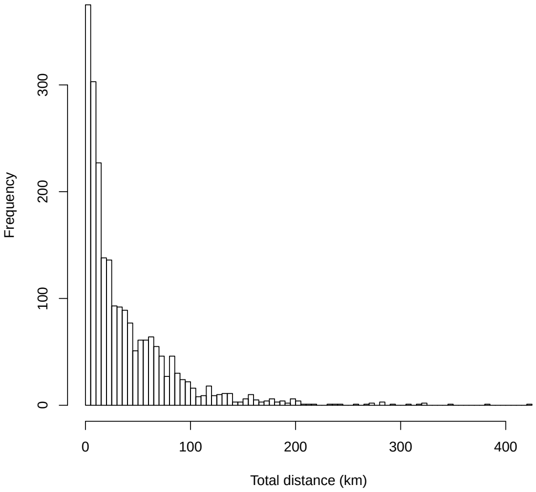

Figure 9: Daily distances on clear days

```
cast (daily, gps_weekday ~ rainyday,
      fun.aggregate=mean, value=’dailydistance’)

  gps_weekday     FALSE     TRUE
    Monday     31.43842  44.35920
    Tuesday    33.01699  34.96554
    Wednesday  34.44736  41.70682
    Thursday   37.21042  48.45119
    Friday     40.92975  40.74351
    Saturday   44.54265  67.99861
    Sunday     51.10177  51.49553
```

Moreover, precipitation makes people move further from home. On days without rain or snowfall, the median of distance from home was 3.8 km. On days with rain or snowfall it was 8.2 km (t-test p-value rain/no rain on distance from home 2.2e-16).

The data seems to indicate that on rainy/snowy weekends people travel further than they do on clear weekends. For simple analysis, we construct two data sets: RainyWeekendAway (Saturday or Sunday, distance from home > 10km, daily rain > 0) and ClearWeekendAway (daily precipitation = 0). The popular destinations among those are indeed different. Ski holiday sites (Leysin, Zermatt) feature in the top-10 addresses of clear weekends, but the popular addresses of rainy/snowy weekends are not ski holiday sites. Rather, they appear to be private residences in cities and towns. Maybe people are visiting their friends and relatives.

Evidently, people visit Leysin and Zermatt during winter months more than in the spring, but the effect is not as strong as with clear day vs. rainy day. The following table gives a summary of mean daily distances in different months, with clear/rainy criteria.

```
cast (daily, gps_year_month ~ rainyday,
      margins=TRUE,fun.aggregate=mean,
      value='dailydistance')

         month    clear  rainy   (all)
1        2009-11  37.03  41.69  38.31
2        2009-12  38.17  51.00  40.99
3        2010-01  36.12  41.18  36.92
4        2010-02  38.46  45.12  40.07
5        2010-03  38.52  53.61  40.62
6        2010-04  40.38  53.64  40.85
7        2010-05  24.58    -    24.58
8          (all)  38.24  46.71  39.67
```

Since Switzerland has many mountains, we continue the analysis by testing whether the weather affects altitude. We test this by grouping the data based on weekdays and months. It seems that during weekends people seem to go to mountains more frequently if the weather is clear. The same phenomenon can be seem during the ski season.

```
    either  rain   clear
Mon  654.2   598.1  620.4
Tue  629.6   579.4  613.2
Wed  592.7   531.2  575.4
Thu  634.1   535.6  601.0
Fri  633.7   543.5  614.7
Sat  805.8   664.1  777.9
Sun  784.7   587.9  768.7
```

```
           either   rain    clear
2009-11    588.45  547.21  577.75
2009-12    672.96  584.69  588.85
2010-01    770.23  593.98  747.44
2010-02    778.65  602.61  739.17
2010-03    600.78  562.98  592.43
2010-04    638.72  575.31  639.21
2010-05    827.77    -     827.77
```

Contrary to our initial hypothesis, we could not find any patterns in visits to a large furniture store. This is quite easy to identify in the data because the store lies in an area where people normally go for no other purpose than to visit it. We could isolate 10 obvious visits in the store in the data (the coordinates remained in the area at least 30 min, direction changes often, speed low) and about 5 less obvious (the person might be stuck in a traffic jam close to the store). Half the obvious visits took place on Saturdays, but there was no correlation with rain.

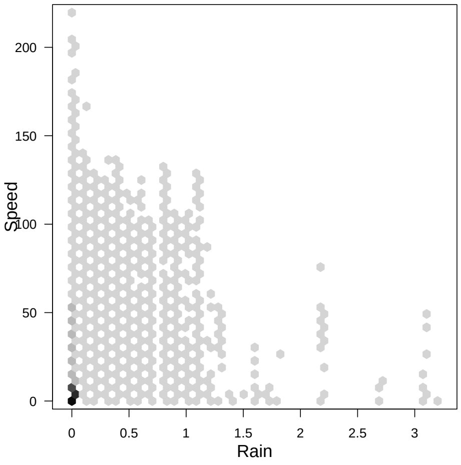

Figure 10: Effect of rain on speed

Next, we studied how precipitation affects speed. This can be done by using:

```
cast (persons, . ~rain, fun.aggregate=max, value='speed')
```

This is shown in Figure 10. The X axis is the precipitation during a 10-minute slot. The p-value of t-test rain/norain on speed is $2.1 * 10^{-16}$

Finally, we classify observations to three classes based on the temperature: cold ($\leq 0\,°C$), cool (> 0 and $\leq 15\,°C$), and warm ($> 15\,°C$). Using this classification, we can test the effect of precipitation and temperature together trying to find out if persons move differently on Sundays than on Mondays.

We can find out that on clear Sundays people move always more than on clear Mondays. On warm clear days people move less than on cool or cold days. Strangely the distance is much longer on Mondays when the temperature is cold and weather rainy or snowy than on clear Mondays.

```
cast (daily, wcat ~gps_weekday ~rainyday,
      fun.aggregate=mean, value='dailydistance',
      subset=gps_weekday == 'Monday' |
             gps_weekday=='Sunday')

rainyday = FALSE
      gps_weekday
wcat     Monday    Sunday
  cold 29.34524 53.714675
  cool 32.90024 53.996604
  warm 24.72435 36.237457

rainyday = TRUE
      gps_weekday
wcat     Monday  Sunday
  cold 54.46814 22.8459
  cool 43.34831 57.1131
  warm      NaN     NaN
```

4. CONCLUSIONS

In this paper we have presented a method for integrating data from different sources for analysis. The mobile data challenge sample, together with weather and address information was used as an example. By integrating the data, we could study how the weather affects people's route selections.

We use the popular and concise RDF formalism to present the data. This has the following benefits: (i) data can be mainly integrated based on its meaning, not the structure, (ii) the validity of the data can be verified by an RDF validator, and (iii) we can use an RDF query language to select a subset of the data, like combining weather and mobility data based on dates. This will facilitate the next step, namely loading the selected data into an analysis tool. In this paper, we have used the R[12] statistics system for analysis. In future work, we also plan to use OLAP database tools. This will help us (i) enhance OLAP with statistical analysis power of R. We will also adapt our earlier research in order to have a sound model for calculating aggregates (like "daily rain in April") [6] and (ii) measure OLAP cube construction times with real data.

5. REFERENCES

[1] E. Codd, S. Codd, and C. Salley. Providing OLAP to user-analysts: An IT Mandate. Technical report, Hyperion, 1993.

[2] F. F. Manola and E. Miller. RDF primer, W3C recommendation 10 February 2004. Technical report, W3C, 2004. (eds). Available at `http://www.w3.org/TR/rdf-primer`.

[3] M. Gonzales, C. Hidalgo, and A-L. Barabasi. Understanding individual human mobility patterns. *Nature*, (453):779–782, 2008.

[4] B. McBride. The resource description framework (RDF) and its vocabulary description language RDFS. In S. Staab and R. Studer, editors, *Handbook on Ontologies*. Springer, 2004.

[5] E. Miluzzo et al. Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application. In *Proceedings of the 6th ACM conference on Embedded network sensor systems*, Raleigh, NC, USA, Nov 2008.

[6] T. Niemi and M. Niinimäki. Ontologies and summarizability in OLAP. In *Proceedings of the Semantic Web and Applications (SWA), A Technical Track of the 25th Annual ACM Symposium on Applied Computing, Sierre, Switzerland*, March 2010.

[7] T. Niemi, S. Toivonen, M. Niinimäki, and J. Nummenmaa. Ontologies with Semantic Web/grid in data integration for OLAP. *International Journal on Semantic Web and Information Systems, Special Issue on Semantic Web and Data Warehousing*, 3(4), 2007.

[8] M. Niinimäki and T. Niemi. An ETL process for OLAP using RDF/OWL ontologies. *LNCS Journal on Data Semantics XIII, Special Issue on "Semantic Data Warehouses'*, 5530:97–119, 2009.

[9] A. Noulas, S. Scellato, and C. Mascolo. An empirical study of geographic user activity patterns in foursquare. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain, July 2011.

[10] C. Ratti, R. M. Pulselli, S. Williams, and D. Frenchman. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, (33):727–748, 2006.

[11] H. Uematsu et al. Balog : Location-based information aggregation system. In *Proc. 3rd International Semantic Web Conference (ISWC2004)*, Hiroshima, Japan, November 2004.

[12] John Verzani. *Using R for Introductory Statistics.* Chapman & Hall/CRC, 2005.