Statistical Analysis and Visualization Services for Spatially Integrated Social Science Datasets

Irfan Azeezullah, Friska Pambudi, Tung-Kai Shyy, Imran Azeezullah, Nigel Ward, Jane Hunter

eResearch Lab, School of ITEE
The University of Queensland

Robert J Stimson
Australian Urban Research Infrastructure Network
The University of Melbourne
overview

1. spatially integrated social science
2. system objectives
3. system overview
4. evaluation
5. future directions
background: spatially integrated social science

Much social science data has an associated geographic location
  - e.g. survey respondent’s location

Use geography to
  - integrate heterogeneous data
  - visualise data
  - perform statistical analysis

Statistical Analysis and Visualization Services for Spatially Integrated Social Science Datasets
IEEE eScience, Chicago, October 2012
background: voting & demographic data analysis

Taylor 2003

- strong correlation between voting for Australian Greens Party and individuals with tertiary education & secularity

Stimson, Chetri, Shyy 2007

- voter support for Australian Labour Party leaned toward asset-poor, multicultural areas
- National Party & One Nation Party competed for votes in asset-rich, monocultural, low income & low education areas
- accurately identified heartlands and marginal areas of voter support for key parties in the next election
background: researcher needs

data assistance
• time consuming to collect
• in multiple formats

analysis assistance
• aligning data with geographies
• coding spatial statistical algorithms

tools exist but are not aimed at social scientists
• complicated
• proprietary
• offline (making repeatability difficult)
• not enough statistical tools to prove hypotheses
objectives

1. web-based system
   • data access & correlation
   • statistical analysis & visualisation
   • geospatial analysis & visualisation

2. Viable system
   • enough data & tools to prove hypotheses

3. Flexible & extensible system
   • incorporate new tools and new data as they become available
system: demonstration

“Find correlations between demographic / socio-economic data and voting patterns during the last Australian federal election”

http://esocialscience.org
Percent primary vote Coalition parties vs. Labor
Melbourne Metro

Data Classification

- 2010 Federal Election Poll Places
- % Primary vote for Coalition vs Labor
- Voting comparison classification:
  - Coalition = Labor
  - Coalition > Labor
  - Labor > Coalition
  - null
Percent primary votes for Coalition overlaid on percent of Anglican (Melbourne)
Regression line fitting for % primary vote for the Coalition parties to the socio-economic variable % of Anglican (Victoria)
Percent primary votes for Coalition overlaid on percent of Generation Y (Melbourne)
Regression line fitting for % primary vote for the Coalition parties to the demographic variable % of Generation Y (Victoria)
system overview
system: data

Create variables

Associate variables with geographies

Group & describe the variables

Levels of geography
1. Polling booth catchments (UQ GPEM)
2. Statistical Local Areas (ABS)
3. Local Government Areas (ABS)
4. Urban Centres & Localities (ABS)
5. Functional Economic Region definitions (CofFEE at Newcastle Uni)

Australian Bureau of Statistics
2006 Australian Census of Population & Housing

Demographic data
Socio-economic data

Australian Electoral Commission voting results
2010 federal election voting results
Polling booth locations

2010 federal election voting results
Polling booth locations

Demographic data
Socio-economic data

Create variables

Associate variables with geographies

Group & describe the variables

Levels of geography
1. Polling booth catchments (UQ GPEM)
2. Statistical Local Areas (ABS)
3. Local Government Areas (ABS)
4. Urban Centres & Localities (ABS)
5. Functional Economic Region definitions (CofFEE at Newcastle Uni)

Statistical Analysis and Visualization Services for Spatially Integrated Social Science Datasets
IEEE eScience, Chicago, October 2012
system: architecture

Statistical Analysis and Visualization Services for Spatially Integrated Social Science Datasets
IEEE eScience, Chicago, October 2012
evaluation:
early researcher feedback

Add
• scatter plots + linear regression graph
• gray scale maps (for papers)

Improve speed
• parallelised multiple regression algorithm
• ported classifier from Java to R

Missing variables on some panes
• now (mostly) create UI from metadata
• auto generate and test for variables in UI
evaluation:
early researcher feedback

What’s up with these polling booth catchments?

<table>
<thead>
<tr>
<th>Polling booth id</th>
<th>Population</th>
<th>Votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>7828</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11913</td>
<td>19</td>
<td>54</td>
</tr>
<tr>
<td>1773</td>
<td>1699</td>
<td>0</td>
</tr>
</tbody>
</table>

7828 – National park ranger’s office
11913 – Perth domestic airport
1773 – Incorrect polling booth location data
evaluation: system

Open source re-implementation provided re-factoring opportunities …

service-oriented architecture
  – open source Geostack
  – Rserve

separated content from presentation
  – previously: graphs generated in R
  – now: R -> XML -> Processing(.js)
    – Graphs
    – Interactive graphs
    – PDF
evaluation: data

Data integration, variable extraction and geography association

• manual process
• still uses proprietary software

Rights

• currently have open access system rather than open data system …
• original census and voting data is freely available
  – can derivative works be re-distributed?
• some of the region definitions contain significant IP developed by higher education institutions
  – still to convince them to share the region definitions
objectives (reflection)

1. **web-based system**
   - data access & correlation
   - statistical analysis & visualisation
   - geospatial analysis & visualisation

2. **viable system**
   - enough data & tools to prove hypotheses

3. **flexible & extensible system**
   - incorporate new tools and new data as they become available
Moving forward

Now supporting the work ...

Integrating our R analysis and Processing visualisation tools into their system

- refactoring to their plugin architecture (OMS3)

Adding new data + analysis tools

- based on researcher focus groups
- beyond demography / voting data
  - transport, health

Automate derived data generation
summary

1. web-based system for spatial-social science
   - data access & correlation
   - statistical analysis & visualisation
   - geospatial analysis & visualisation

2. aimed for viable system
   - enough data & tools to prove hypotheses

3. aimed for flexible & extensible
   - incorporate new tools as they become available
Acknowledgements

This project is supported by the Australian National Data Service (ANDS) and the Australian Urban Research Infrastructure Network

ANDS and AURIN are supported by the Australian Government through the National Collaborative Research Infrastructure Strategy Program and the Education Investment Fund (EIF) Super Science Initiative.
QUESTIONS?

http://esocialscience.org/

n.ward4@uq.edu.au