Andy Turner On **MoSeS**

Andy Turner

http://
www.geog.leeds.ac.uk/people/a.turner/

Overview

- A General Introduction to the Aims and Objectives of MoSeS and an Outline of Current Work
- A Personal Perspective of Collaborative Working in e-Social Science Based on My Experience
- Describe how the demographic model of the UK we are developing might be used in policy analysis

Outline

- Introduction
  - What is MoSeS?
  - Who are we?
- MoSeS
  - Aims and Objectives
  - Demographic Modelling
  - Demonstration Portal
  - Collaborative Links and other Projects

Introduction

What is MoSeS?

- Modelling and Simulation for e-Social Science
  - http://www.ncess.ac.uk/research/nodes/MoSeS/
  - e-Social Science being the application of e-Science concepts to social science problem domains
    - e-Science is enhanced science that uses the Internet, software tools and structured information for collaborative work
- MoSeS is a Node of NCeSS
  - Part of a UK collaborative partnership developing e-Social Science
  - The key part of it's program of work is to develop an individually based demographic model of the UK for 2001 to 2031

Who am I and Why am I here?

- I am based in the School of Geography at the University of Leeds
- My background
  - Mathematics and Geographical Information Systems
  - Computational Geography research since 1997
  - Data Analysis and Modelling
  - Java Programmer
- I am part of various research organisations
- An e-Research enthusiast
  - I first heard about e-Science 2 years ago
- Since beginning work on MoSeS
  - An example of e-Social Science in action
    - e-Social Science is a way of working
      - I am trying to employ and develop that
  - Open Research Collaborator
- Semantic Web Enthusiat
- Part of MoSeS and NCeSS
- I am here to communicate, meet people, make friends and encourage collaboration
- http://www.geog,.leeds.ac.uk/people/a.turner/

Who are you and why are you here?

- Social Scientists?
  - e-Social Scientists?
- e-Social Science Sceptics
  - To be won over…
  - Not to be antagonised!
- Unengaged
  - To be engaged…
- Enthusiast and Practitioners
  - To be encouraged…
  - To consider collaboration…
- To half listen for something interesting whilst getting on with some work
- Whoever you are you are welcome!


E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL


UNIVERSITY OF
LEEDS


CCG
Centre for Computational Geography


NCeSS
National Centre for eSocial Science

MoSeS

Aims and Objectives

- Raise awareness of eScience and eResearch
- Develop practical geographical e-Social Science applications demonstrating the potential of Grid Computing
- Model the UK human population at individual and higher organisational levels
  - households, communities, regions
  - disparate and/or geographically diffuse organisations and society
  - service orientated government
- Develop and package a suit of modelling tools which allows specific research and policy questions to be addressed with demonstrator applications for:
  - Health
  - Business
  - Transport

Initial Tasks

- Develop methods to generate individual human population data for the UK from 2001 UK human population census data
- Develop a Toy Model
  - Dynamic agent based microsimulation modelling toolkit and apply it to simulate change in the UK
- Develop applications for
  - Health
  - Business
  - Transport

Challenges

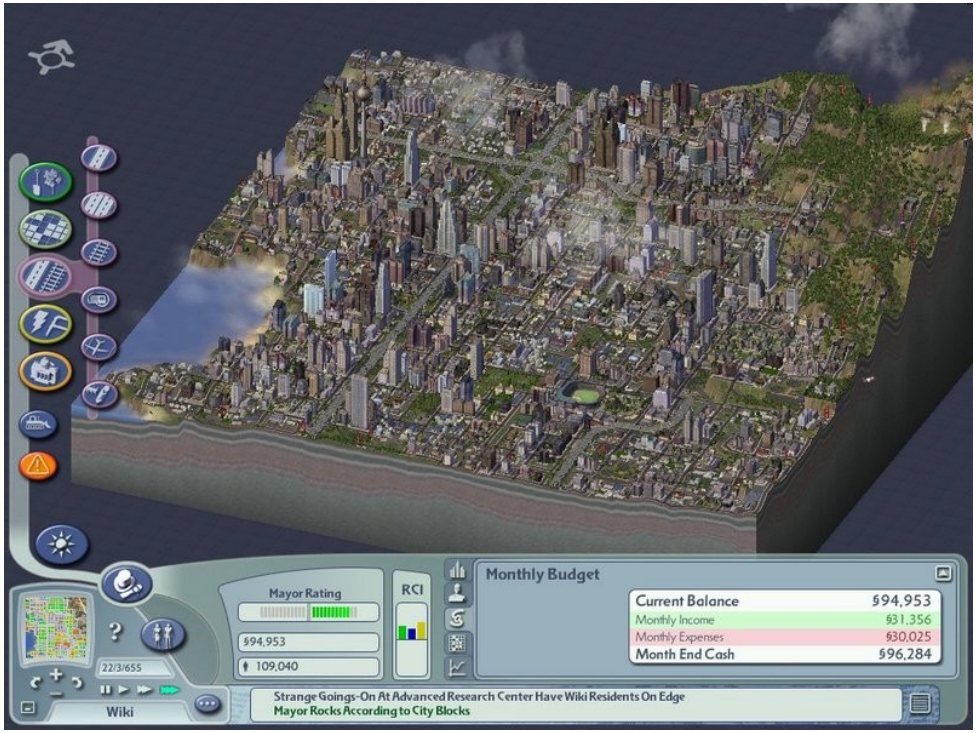
- Grid enabling the data and tools
- Visualisation
  - Google Earth
  - Computer Games
- Collaboration
- Retaining a problem focus
- Design and Development

Generic MoSeS Approach

- MoSeS to date has approached Modelling and Simulation from a specific angle
  - Geographic
  - Demographic
  - Contemporary
  - About the UK
  - Targeted towards supporting a developing set of applications
- It is not a requirement to make it clear what steps can be followed by other Social Scientists wanting to Model and Simulate something different
  - However, the generic work of MoSeS should be relevant and we are working towards this

MoSeS Vision

- Suppose that computational power and data storage were not an issue what would you build?
  - SimCity
    - http://en.wikipedia.org/
    - For real on a national scale



MoSeS Rationale

- The idea is to provide planners, policy makers and the public with a tool to help them analyse the potential impacts and the likely effect of planning and policy changes.
- Example Application:
  - There may be a housing policy to do with joint ownership, taxation and planning restriction legislation that can be developed to alleviate problems to do with lack of affordable housing and workers without precipitating a crash in the housing market and economy as a whole
  - A balanced policy may be easier to develop by running a large number of simulations within a system like SimCity for real to understand the sensitivities involved

MoSeS First Steps

- The development of a national demographic model
- The development of 3 applications
  - Health care
  - Transport
  - Business
- The development of a portal interface to support the development and resulting applications by providing access to the data, models and simulations and presenting information to users (application developers) in a secure way

The development of a national demographic model

Required Characteristics

- Individual level
  - So that people can be modelled as individual agents
    - Individuals to be grouped into households
- Dynamic
  - For the period 2001 to 2031 based on an annual time step

- Based on 2001 UK Human Population Census Data
  - Individual and Household SARS
  - Census Aggregate Statistics (CAS)
    - Area Based
      - OA, Ward, LAD
- Enriched with Variables from other Data added by Microsimulation using probability distributions and variables in these data that are understood to map onto census variables

Division of Labour

- I set to develop the base population for the year 2001
  - Population initialisation
- Belinda Wu set to develop a dynamic simulation model with the processes of birth death and migration explicitly modelled
  - Dynamic simulation

Population initialisation

- http://www.geog.leeds.ac.uk/people/a.turner/pr
- Essentially the task is to select from:
  - The 3% Individual Sample of Annonymised Records (ISAR) for the UK
    - 1843525 Individual Records
    - Used initially for all population and laterly for Communal Establishment Population
  - The 1% Household Sample of Annonymised Records (HSAR) for the England and Wales
    - 225436 Households 525715 Individual Records
    - Used later when it became available for Household Population

- Select a set of records with aggregate statistics that are a good match for those in each CAS area
- This can be done in various ways
  - The approaches we are trying include
    - Iterative Proportional Sampling
    - Genetic Algortihm:

Permutations

- Given the population (p) of an Area we want to select a sub-sample of this size from the number of records in the ISAR n
- The general formula for finding the number of permutations of size p taken from n objects npPermutations is:

$$\text{npPermutations} = \frac{n!}{(n-p)!}$$

- Approximately n^p

Computation

- Number of potential solutions too great to find the best fitting solution by a brute force search
  - The number of potential solutions is even greater for larger regions (although there is a small consolation that there are less of them!)
- Fortunately we are only interested in specific types of solution and can constrain the search
- For some criteria hard constraints are appropriate and for other variables, optimisation is preferred within these constraints

Constraints

- What can we constrain to?
  - There are limits
    - The more detailed the constraint criteria the less likely it can be met
  - The ISAR is only a 3% sample
  - Specific CAS tabulations
    - The aggregations of variables are bespoke
    - Beware of errors especially systematically introduced disclosure control measures
  - Census data are estimates and contain unknown levels of error
- What is most important to ensure is right?
  - Age/Gender profile
  - Number of Household Reference People
  - Household Composition
  - Social Class
  - Health status etc…

CAS

- Key Statistics Tables
  - 31 tabulations
  - E.G KS001
    - Usually Resident Population
    - 6 cells
- Standard offerings
  - 53 cross tabulations
  - E.g. CS001
    - Age/Sex/Resident Type
    - 250 cells
- Themed Tables
  - 6 cross tabulations
  - E.g. CT001
    - Theme Table On All Dependent Children
    - 348 cells
- Univariate Tables
  - 43 tabulations
  - E.g. UV003
    - Sex
    - 3 cells

Constraint and Optimisation using Key Statistics

- As a first step we constrained by age and ensured that we had the correct number of household reference people
  - Makes it easier to construct households for the dynamic model of Belinda (Toy Model)
- Used a Sum of Squared Errors (SSE) fitness function for a number of aggregated variables
  - Measure of the difference between aggregate counts from the ISAR records and the published and aggregated CAS Key Statistics
    - Initial focus on health, household composition and employment status

Progress

- A first Individual level UK population dataset of 58789293 records was produced in January 2006
  - This was based purely on the ISARs
  - Belinda found it difficult to form these into households based on a Household Formation Routine she developed in Toy Model
- A second set was then produced which used Belinda's Household Formation Routine to ensure that the fit of the results were not only good as per the CAS, but also that the right number of the right sorts of households could be formed.

The next step…

- What arguably should have happened then, is that we focus on writing up the method and focus resources on developing the dynamic model
  - For various reasons this did not happen
- Instead the HSARs were released and all the population intialisation work was re-iterated
  - The HSARs were to be used to form Household Population (HP) and the ISARs were used to form Communal Establishment Population (CEP)

Documentation takes over

- By this stage I was developing web pages to organise information about MoSeS for reference purposes
  - The main driver for this was to make our own operation more efficient
  - I felt it was also important to expose as much of what we were doing to others as we could
  - I was documenting work on the population initialisation and trying to leave a trail of meeting notes
  - I made an effort to link all the information together as I felt sure that someone should be doing this, this was something that good e-Social Science would do…

More progress

- Several months later, I produced a new dataset using OA constraints
  - Some visualisations of the results allowed their quality to be scrutinised by the population experts on the team
  - Their quality was not deemed good enough
    - For various reasons it was decided that MSOA constraints must be used
- I automated the process of visualising the results
- A short time later I discovered that the control constraints could not be met with the imposed restriction of Sampling Without Replacement (SWOR)
  - The restriction of SWOR was lifted which required an almost complete reworking of the code

Yet more progress

- Sampling With Replacement (SWR) required less checks and significantly simplified matters in terms of the methods used
  - Focussing on Leeds, the population initialisation has been going through a review process for about 6 months
    - The results are being compared with an Iterative Proportional Sampling method
    - Some criteria for an acceptable goodness of fit are being developed
      - I am hoping that with these I can finally satisfy the population experts with the results from this component of the work

Process is as important as progress

- An upside of having done all this work in more than one way is that the results can be compared
  - Indeed so much can be compared it might take a lifetime to do so!
- If all goes well in the next period a population data set will be produced for Leeds that satisfies the advisors and which is based on HSAR for HP and ISAR for CEP
- I will then focus on generating results for the UK, documentation and enriching the data with additional variables

Dynamic Simulation

- Meanwhile Belinda has been developing a dynamic simulation
  - An annual step change of the population from 2001 to 2031
  - Migration is proving to be no less difficult now than it ever was!
  - This work is never ending
    - A more complex model can be made more realistic and so on infinitum
    - For the time being we are using our experts to constrain the model and use assumptions appropriately as a substitute for an incredibly detailed model of migration
  - Results are coming and Paul is working on the portal to interface for our application developer/users

Demonstration Portal in Action

- http://geo-s12.leeds.ac.uk:8080/grids

Links and Overlaps with other work

- SIM/UK
- NIEeS Grid GIS Working Group
- JISC OGC Grid Collision
- GeoVUE
- NCeSS e-Infrastructure project

Recap

- In MoSeS we are developing a demographic model of the UK
  - Comprising of individual people that occupy the UK environment and move about it through time interacting in numerous ways
  - Each individual will have family, household and social networks and reasonably complex characteristics and behaviour
- We are trying to build a platform (data and tools) for simulating change in the UK
- We are operating in a collaborative way and trying to stand on the shoulders of giants

Acknowledgements

- Thanks to all the MoSeS team in particular Belinda, Mark and Paul who are working on this day to day
- Thanks to NCeSS for being a great bunch of people to work with
- Thanks to all involved in eResearch for improving our hardware, software and data resources so that we can all do our bit to better understand and plan our future