# e-Social Science:
## scaling up social scientific investigations

**Alex Voss, Andy Turner, Rob Procter**
**National Centre for e-Social Science**
**Gabor Terstyanszky, Gabor Szmetanko, Tamas Kiss**
**CPC @ Westminster**

**Presentation at ISGC 2009, Taipei, Taiwan, 2009-04-22.**

# Overview

- **Background**
- **Introduction**
- **MoSeS**
- **GENESIS**
- **Demographic Modelling**
  - Population Reconstruction (Initialisation)
  - Dynamic Simulation
- **Experience**
- **Organisation**
- **Scaling Issues**
- **Future Work**
- **Acknowledgements**

# Background

- **Much social science does not use advanced ICT but emergence of new analytical methods is driven by:**
  - Increased availability of data about social phenomena
  - Issues with data management and integration
  - Challenges to analyse social phenomena at scale
  - Challenges to inform practical policy and decision making (e.g., evidence-based policy making)
- **National Centre for e-Social Science (NCeSS) in the UK is investigating ways to respond to these challenges.**
- **EUAsiaGrid is supporting e-Social Science amongst other application domains**

# Introduction



- **Virtual worlds rich in detail are being developed**
  - Digital representations (of parts) of Earth are being developed
  - Necessarily generalised models
  - Interact with the ‘real’ world
  - Socio-economic
- **There can always be more detail**
  - Higher spatial and temporal resolution
  - More and more detailed attributes
  - Geography and social science is no different to any other type of science in this respect
- **We are all geographers to some extent and we all interact in some way with the object of study**

# MoSeS



- **Modeling and simulation approaches for social science**
- **First phase research node of NCeSS**
- **Core contemporary demographic model of the UK based on UK census data and other datasets**
- **Using agent-based simulation to project population forward in time by 25 years**
- **Explicitly model births, deaths, migration, changes in health status etc…**
- **Applications in transportation research, health and social care planning and business applications**

# GENESIS



- **Uses and builds on MoSeS**
- **A team involving experts in geovisualisation from UCL**
- **Two development strands**
  - Theoretic models based on restriction free data
  - Models seeded with more restricted access data
- **More theoretical**
  - Computational limits
  - Investigating what visualisations are useful
  - Considering how to do validation models
- **Less emphasis on developing specific applications**
  - Applications being considered in transportation planning
  - Respond ad hoc to what is in the public interest
- **Daily activity models**

# Demographic Modelling I



- **Generation of an individual level population data for the UK**
  - Based on 2001 census data
  - Works with ‘public release’ versions of census that are restricted,
    - Census Aggregate Statistics (CAS) at Output Area Level
    - 1% of population (anonymisation)
  - Reconstructed data has same attributes as real population and same number of individuals but is still anonymised
  - Uses a genetic algorithm to select a well fitting set of sample of anonymised records to assign to an output area
  - Need for attributes in the SAR to be matched with those in the CAS
    - This is often complicated because of different categories
      - *Aggregation to a lowest common categorisation*

# Demographic Modelling II



- **Dynamic modelling**
- **Daily activity modelling**
  - Commuting
  - Retail modelling
  - Transportation
- **Population Forecasting**
  - Annual time step
  - Birth
  - Death
  - Migration

# Experiences



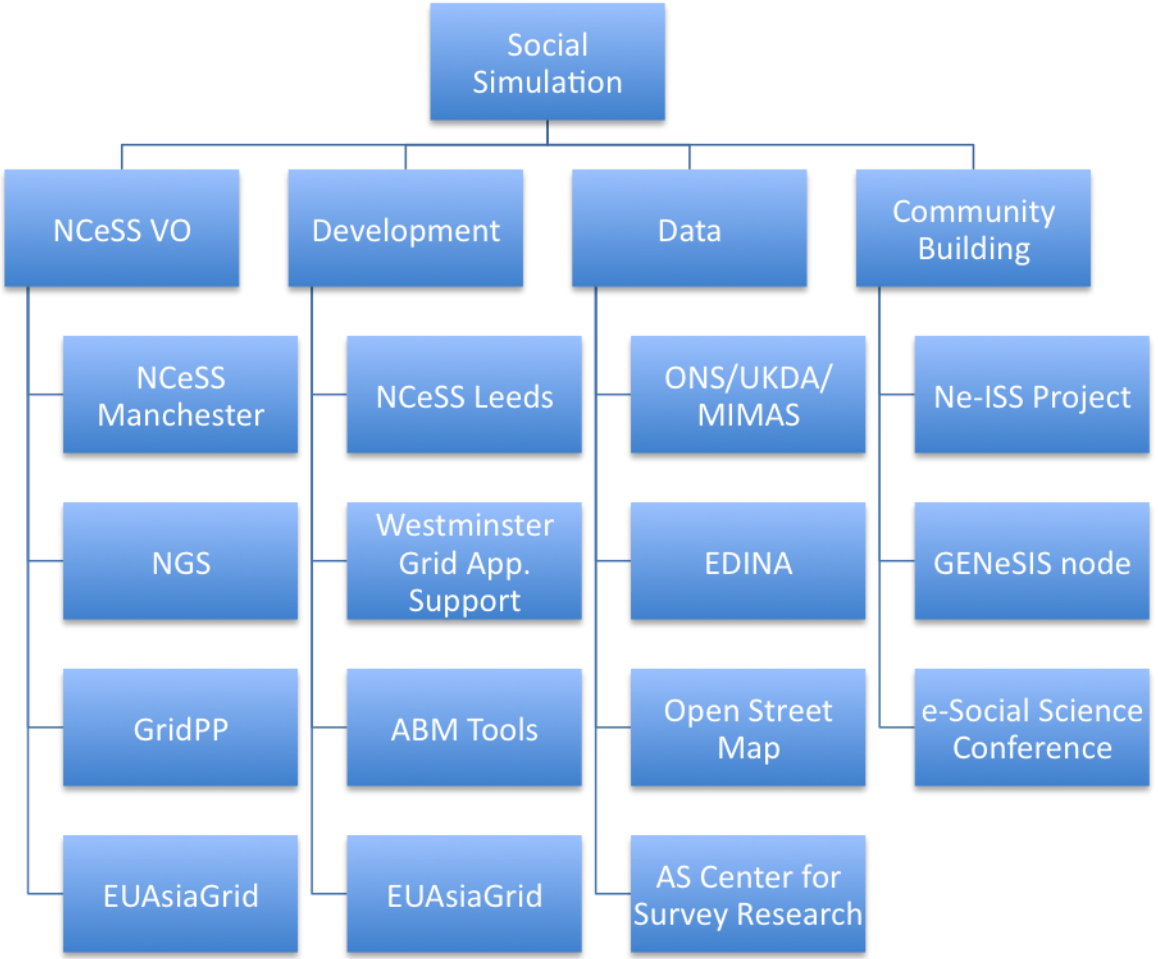
- **Integrating existing code into grid environment required some changes to source code**
  - management of input arguments
  - code scalability
  - log management
  - error handling
- **Finding the right input size and parameters for testing to keep execution times low**
- **Making sense of execution failures**
  - lack of ways to debug code in distributed environments

# Experiences II



- **Step-wise process works well,**
  - ensures we encounter problems piece by piece
  - allows us to comply with data protection / licensing
- **Population reconstruction is resource intensive**
  - may run up against limits on wall clock time
- **Importance of ‘at elbow’ support**
  - but hindered by data protection/licensing issues
- **Licensing means we need to limit execution to UK resources**
- **Setting up VO to support secure sharing of data**

# Organisation



# Scaling Issues I



- **Simulations**
  - Need to find ways to map to different architectures, both HPC and HTC
  - Need to deal with large memory requirements and limitations imposed by OS, JVM and Java libraries
  - Exploring Terracotta
    - Distributing computation
    - Virtual heap space
    - Dependability
    - Advice would be very welcome…

# Scaling Issues II



- **Population model size and sophistication**
  - From town size to country size (and beyond)
  - Number of variables
  - Number of constraints
- **Number of cores used**
  - To reduce runtime
  - Need to go beyond using only one site
- **Community**
  - Open development – needs tool support
  - Number of users – requires hardening of code & documentation

# Future Work



- **Next steps until code runs in Taiwan with Taiwanese data**
  - Proof of concept execution on Quanta cluster at ASGC
  - Definition of data outputs from
- **Develop submission to exploit multiple NGS nodes and EGEE Compute Elements**
- **Improving data and code staging**
- **Moving from population reconstruction to supporting the simulation process**
- **Integration into ‘science gateway’ for the social sciences and developing a repository for models**

# Acknowledgements


- **National Centre for e-Social Science**
  - MoSeS Node: Mark Birkin (PI)
  - GENeSIS Node: Mike Batty (PI)
  - NCeSS Hub: Peter Halfpenny and Rob Procter
- **EUAsiaGrid Consortium**
  - Marco Paganoni (Project Director)
- **CPC at Westminster University**
  - Gabor Szmetanko
  - Gabor Terstyanszky
  - Tamas Kiss
- **GridPP**
  - Jens Jensen and Jeremy Coles
- **National Grid Service**
  - Jason Lander and Shiv Kaushal (Leeds), Steven Young (Oxford), Mike Jones (Manchester)