Developing big data methods in environmental epidemiology

Antonio Gasparrini

London School of Hygiene and Tropical Medicine
United Kingdom

Workshop Big data at LSHTM: who is doing what?
Centre for Statistical Methodology
14 March 2017 – LSHTM

Environmental epidemiology

Definition (Wikipedia): *"the branches of epidemiology concerned with the discovery of the environmental exposures that contribute to or protect against injuries, illnesses, developmental conditions, disabilities, and deaths; and identification of public health and health care actions to manage the risks associated with harmful exposures"*

Peculiarities:

- **Widespread exposure** to environmental factors – *e.g.*, air pollution – often affecting the whole population
- Often **small risks** – *e.g.*, a RR of 1.0051 (95%CI: 1.0007–1.0093) for an increase of $10\mu gr/m^3$ of PM_{10} (Samet NEJM 2000)
- Need to perform epidemiological analyses on **large populations**

Traditional analyses: limitations

- Exposures assigned or reconstructed with **low temporal and/or spatial resolution** – *e.g.*, over large areas using central monitors – with issues such as ecological biases and measurement error
- Health data obtained from **administratively collected** databases, and often aggregated over large areas: lack of individual information, no knowledge on susceptibility factors

'Big data' opportunities

New **big data technologies** are already transforming the landscape of medical and epidemiological research, and offer opportunities also in environmental epidemiology

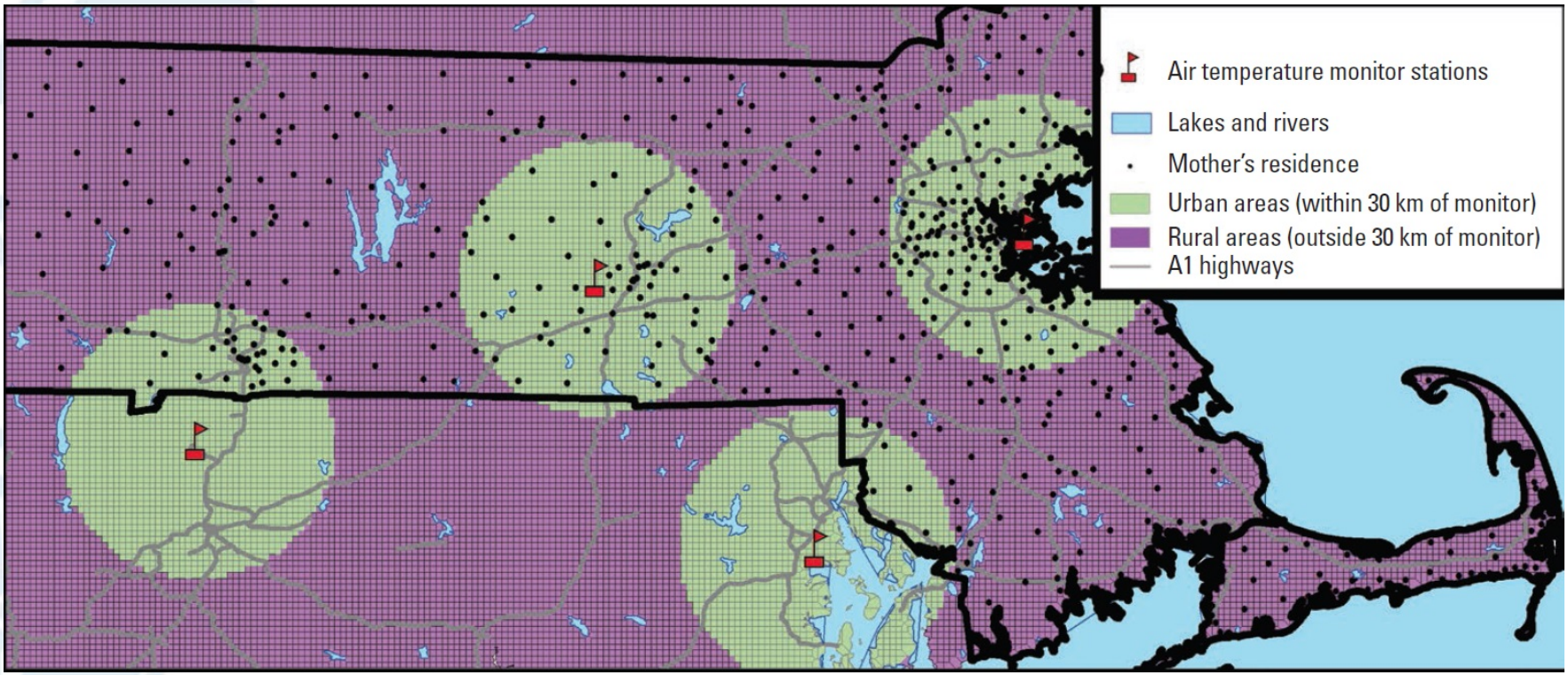
For example:

- New resources providing **high-resolution measurements** of environmental exposures, such as remote sensing data from satellites and emission/dispersion modelling
- Linkage between **electronic health record** databases with detailed information on health outcomes and risk factors on large populations

Chance to move **from aggregated to individual-level** investigations

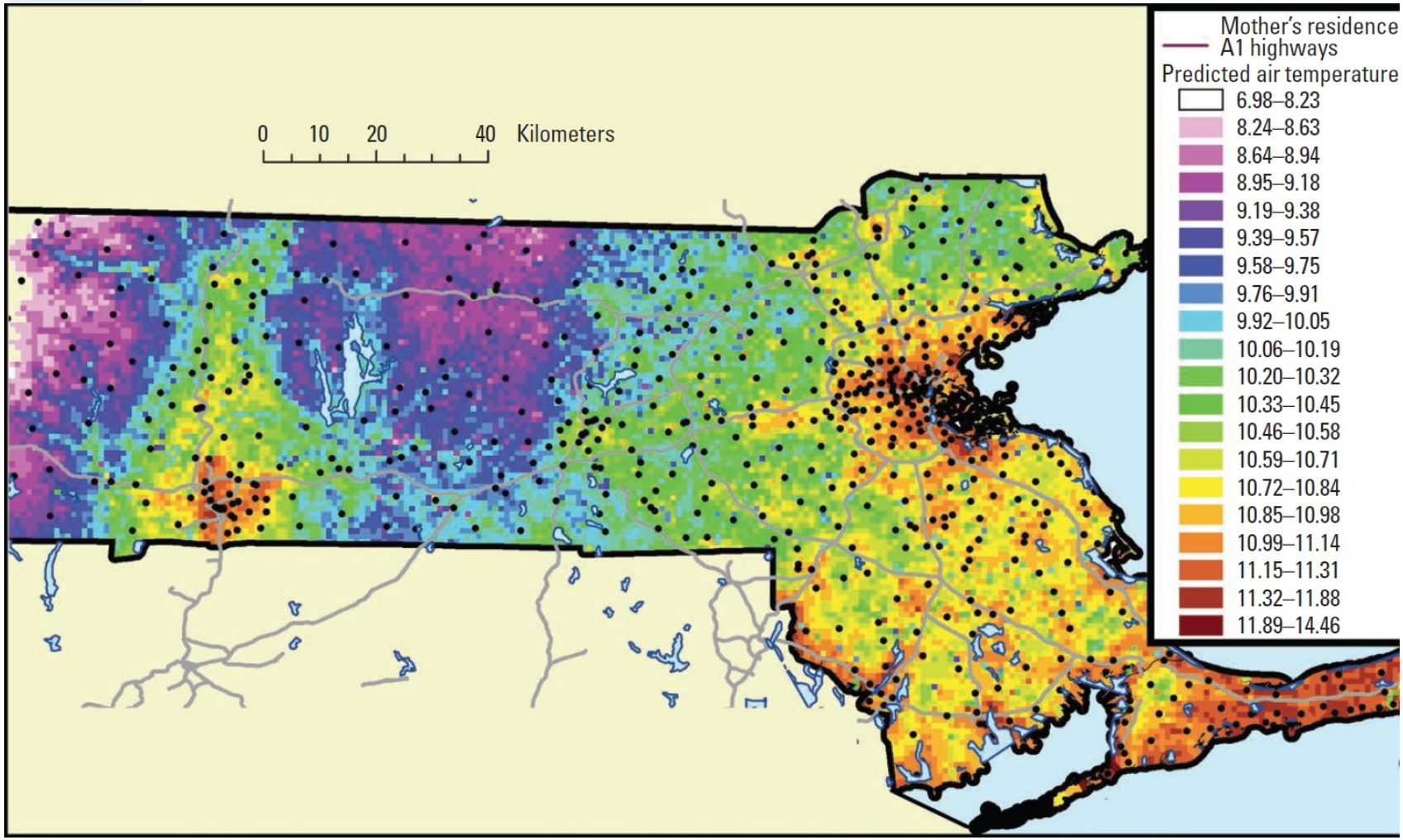
'Traditional' approach

From Kloog EHP 2015



'Big data' approach

From Kloog EHP 2015



Statistical/computational methods

- **Hybrid models** to integrate multiple exposure sources in high-resolution spatio-temporal maps
- **Two-stage** designs to separate the analysis across sub-areas and then pool with meta-analytical methods
- **Self-controlled case-only** designs to restrict the analysis to cases, who act as their own controls (case-crossover, case series)
- **Computational techniques** for reduction and partition of estimation algorithms, ideal for multi-core computation

In summary

- Traditional measurements of environmental exposures can be integrated with newly available sources, such emission/dispersion models and remote sensing data from satellites, to generate exposure maps with high spatial and temporal resolution
- The linkage of existing cohorts with multiple sources of exposure data and administratively collected electronic health records can form rich datasets including large collections of variables on individual characteristics
- This wealth of data can be used to determine individual risk profiles with longitudinal measures on time-varying exposures, health outcomes and susceptibility factors, significantly extending the analytical capability of environmental health studies