Data for a Smart City
Joining the Dots Conference 27th February 2018
Dr Paul Hewson, Senior Data Scientist

http://www.cityscience.com  paul.hewson@cityscience.com
1 Background

2 Road Injury Risk

3 Predictive Modelling
Cities

- 1950: 746 million people lived in cities
- 2014: 3.9 billion people lived in cities
- 2050: Over 2 in 3 people will live in cities (estimated)

This matters to (a) cities and (b) the whole world.

---

\(^1\)Source: UN (2014) *World Urbanization Prospects* UN DESA Population Division
Cities

- 1950: 746 million people lived in cities
- 2014: 3.9 billion people lived in cities
- 2050: Over 2 in 3 people will live in cities (estimated)

This matters to (a) cities and (b) the whole world.

- 70 cities are bigger than Wales,

---

¹Source: UN (2014) *World Urbanization Prospects* UN DESA Population Division
Cities

- 1950: 746 million people lived in cities
- 2014: 3.9 billion people lived in cities
- 2050: Over 2 in 3 people will live in cities (estimated)

This matters to (a) cities and (b) the whole world.

- 70 cities are bigger than Wales, 38 bigger than Scotland.
- 37 are bigger (but not better) than Yorkshire.

It's cities, not administrative entities that have to compete globally.

---

1 Source: UN (2014) *World Urbanization Prospects* UN DESA Population Division
Optimising Cities

- 4000 vehicles
- Most efficient solution?

Diagram:

- Start: O
- Path: O → b → D → c → O
- Numbers:
  - 45
  - x/100

Equation:

\[
\text{optimal travel cost} = \frac{x}{100} + \frac{45}{5} + \frac{45}{2} \text{ vehicles}
\]
Optimising Cities

- 4000 vehicles
- Most efficient solution?
  - 2000 travel $O \rightarrow c \rightarrow D$,  
  - 2000 travel $O \rightarrow b \rightarrow D$
- \[ \therefore \text{optimal travel cost is} \quad 65 = 45 + \frac{2000}{100} \quad \text{(or} \quad \frac{2000}{100} + 45) \]
Optimising Cities

Optimal travel cost:

\[
\text{8} = \text{4000 vehicles} + \text{Most efficient solution?}
\]
Optimising Cities

- 4000 vehicles
- Most efficient solution? 4000 $O \rightarrow b \rightarrow c \rightarrow D$,
- Optimal travel cost:
  
$$80 = \frac{4000}{100} + \frac{4000}{100}$$
- If *you* switch routes, $\frac{4000}{100} + 45$ or $45 + \frac{4000}{100} = 85$
The “Price of Anarchy”

- This *specific* phenomena does exist $^2$. Is it very common?
- We $^3$ individually pick the best route for ourselves.
- Our individual choice may make a city more congested than if we co-operated with a benevolent routing dictator
- This is just one example of how traffic on road networks behaves in very non-intuitive ways

---

$^2$e.g., Cheonggyecheon, South Korea and 42nd Street, New York 1990  
$^3$according to economic theory
Speed and flow models
Speed flow relationships

- Flow models: traditional, well validated (Census, Loops and Counts etc., Mobile Phone)
- Speed models: machine learning + various (GPS tracker vehicles)
- Allows us:
  - Assess performance of road
  - Assess “stress”
Speed and flow and road injury

- Compute collisions / flow for most roads in a City per time period
- Fit point process models to investigate the effect of actual speed
- Look for gross anomalies ... and assess the strength of evidence to claim them as such
- Generalise / predict elsewhere

The more we condition on geographic and sociodemographic factors the more we can generalise the model. Generalisable models that make out of area predictions are testable
Geographic and demographic

4d plot collisions / predictors

Hilliness level
Implications

Average cyclist: 30 minutes

A measure of “porosity”

www.cityscience.com/blog/analytics-for-healthy-streets-part-1
Background

Road Injury Risk

Predictive Modelling

Junctions 1: crash profile

Not an exact science (but what is?)

- For each crash, assess type
- We also assign a reliability score
- Build predictive models for junction profiles on reliability weighted crash data

\[1\text{e.g. M. Stone, J. Broughton (2003) “Getting off your bike: Cycling accidents in Great Britain in 1990-1999” Accident Analysis & Prevention 35:549-56}\]
Junctions 2: flow profiles

- Junction profiles are complex
  - How many people are making opposed right turns?
  - How much opposing traffic exists?
  - What are the demographics of the road users

- We can estimate association between injury risk\(^a\) and junction profiles

- Are some “routes” higher risk than others?

\(^a\)Have to use probabilistic models
Data Fusion

- Data = (1, 3, 2, 4, 5).
- Arithmetic mean = \( \frac{1}{5} (1 + 3 + 2 + 4 + 5) = 3 \)

---

Data Fusion

- Data = (1, 3, 2, 4, 5).
- Arithmetic mean = \( \frac{1}{5}(1 + 3 + 2 + 4 + 5) = 3 \)
- Or arithmetic mean = \( \frac{1}{5}1 + \frac{1}{5}3 + \frac{1}{5}2 + \frac{1}{5}4 + \frac{1}{5}5 = 3 \)

---

Data Fusion

- Data = (1, 3, 2, 4, 5).
- Arithmetic mean = \( \frac{1}{5}(1 + 3 + 2 + 4 + 5) = 3 \)
- Or arithmetic mean = \( \frac{1}{5}1 + \frac{1}{5}3 + \frac{1}{5}2 + \frac{1}{5}4 + \frac{1}{5}5 \) = 3
- Why not use \( \frac{1}{20}1 + \frac{2}{20}3 + \frac{4}{20}2 + \frac{5}{20}4 + \frac{8}{20}5 \) = 3.75

(Almost) a.k.a. **Exponentially Weighted Moving Average**, used routinely!\(^4\)

---

Data Fusion

- Data = (1, 3, 2, 4, 5).
- Arithmetic mean = \( \frac{1}{5}(1 + 3 + 2 + 4 + 5) = 3 \)
- Or arithmetic mean = \( \left( \frac{1}{5}1 + \frac{1}{5}3 + \frac{1}{5}2 + \frac{1}{5}4 + \frac{1}{5}5 \right) = 3 \)
- Why not use \( \left( \frac{1}{20}1 + \frac{2}{20}3 + \frac{4}{20}2 + \frac{5}{20}4 + \frac{8}{20}5 \right) = 3.75 \)

(Almost) a.k.a. Exponentially Weighted Moving Average, used routinely!\(^4\)

- Or instead of recency, give more weight to data you trust more (this idea is as old as Gauss)
- Or give weight to data that is more important
- Sometimes data cannot be compared directly, but we can imagine that it is correlated with a “latent” variable (psychologists do this all the time)

A lot of data exists already but are not used

Why don’t more UK cities have detailed crash rate maps?

Pollution = $f(\text{Flow/Speed})$; \ldots we already have pollution models! (and we have pretty maps of them!)

Is there an unwillingness to use statistical models to do the fusion?

- No data are perfect
- Statistical models let you carry uncertainty from one dataset to another

Doing data fusion is very powerful - don’t necessarily need super-trendy datasets to add value!