Exploring the relationship between Strava cyclists and all cyclists*.

Dr. David McArthur, Dr. Jinyun Hong, Dr. Mark Livingston, Kirstie English

*This presentation contains preliminary results which are subject to change. Please do not cite.
Active travel

• Walking and cycling can generate large benefits
  • Reduced congestion
  • Reduced emissions
  • Improve health
  • Time savings

• Transport Scotland wants 10% of journeys to be made by bicycle by 2020; with cities responsible for achieving this
Do interventions work?

- Evaluating the effectiveness of interventions is difficult due to the lack of data
- Manual counts take place on specific links/points, but these are expensive and hence infrequent
- Automatic counters can be used but these are also expensive and tend to be sparsely located
- Maintenance and calibration is required to keep them working properly
New data

- Activity tracking apps are used by many people and provide valuable new data about activities
- The Strava cycling app uses GPS to track cyclists’ journeys
- This offers the possibility of having data at a fine spatial and temporal scale for a large number of people
- The data are already being collected all over the world
• The name is taken from the Swedish word sträva, meaning to strive
• It can be used to track running and cycling activities
• Users can track their activities over time and compare to the activities of their friends or the user community
• Users can also compete in competitions
• The app comes in a free and premium version. The premium version offers extra features and costs £5.99 a month or £49.99 per year
• Users have to start and stop the tracking
• They can tag whether or not their trip is a commute or not
• Strava also gather some demographic information about their users
The movement data collected by the app is raw GPS trajectories represented as a triple (latitude, longitude, timestamp)
The GPS trajectories are not made available to researchers. The data is aggregated and provided to researchers/planners through Strava Metro. Data are provided as:
- Origins and destinations with route information (at output area level)
- Minute-by-minute link counts of cycling flows
- Information about waiting times at junctions
- Aggregate demographic information
Problems

- We know not all cyclists use the app for every journey
- It is unlikely that a random sample of cyclists use the Strava app
- In Glasgow in 2015 there were 13,684 athletes who recorded 287,833 activities
- The median distance was 14.9 km
- Of this sample 11,216 were male (1,698 female)
- Can the sample tell us anything useful?
Our approach

• Firstly, we can visualise the data and do a basic sanity check; the patterns look like what we would expect given our knowledge of Glasgow
• We can compare it to other sources of data
• We use the annual two-day cordon counts which are conducted in Glasgow city centre
• We match the links where the counts take place to the same link in the Strava data
Cordon Count (CC)

- Cycle trips are counted in blocks of 30 minutes for 14 hours over two days in September each year.
- We use data from 2013, 2014 and 2015.
- We aggregate both the CC and the Strava data into four different temporal scales, specifically by:
  - Hour
  - Commuting time (peak hours versus non-peak)
  - Day
  - Two-day (i.e. annual)
## Correlations

<table>
<thead>
<tr>
<th></th>
<th>Sample size</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hourly</td>
<td>3192</td>
<td>0.781</td>
</tr>
<tr>
<td>Peak Vs Non-peak</td>
<td>684</td>
<td>0.861</td>
</tr>
<tr>
<td>One day</td>
<td>228</td>
<td>0.882</td>
</tr>
<tr>
<td>Two days</td>
<td>114</td>
<td>0.887</td>
</tr>
</tbody>
</table>
Further work

• We have some additional hypothesis about how these correlations may vary:
  • Does Strava have a higher market share in rich areas e.g. the West End of Glasgow?
  • Is the market share of Strava changing over time?
  • Does the weather affect the percentage of cyclists using Strava?
  • Does the time of day affect the share of cyclists using Strava?
Models

- We have experimented with negative binomial regression models.
- The number of total cyclists is modelled as a function of, among other things, the number of Strava cyclists.
- This allows us to explore the factors influencing the link between the cycling flows.
- It also allows us to adjust the Strava flows to an estimate of total flows across the network.
<table>
<thead>
<tr>
<th>Independent</th>
<th>Model 1</th>
<th></th>
<th>Model 2</th>
<th></th>
<th>Model 3</th>
<th></th>
<th>Model 4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>P=</td>
<td>Coefficient</td>
<td>P=</td>
<td>Coefficient</td>
<td>P=</td>
<td>Coefficient</td>
<td>P=</td>
</tr>
<tr>
<td>Strava</td>
<td>0.084</td>
<td>0.000***</td>
<td>0.265</td>
<td>0.000***</td>
<td>0.098</td>
<td>0.000***</td>
<td>0.105</td>
<td>0.000***</td>
</tr>
<tr>
<td>Commuting (ref non-commuting)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td>0.317</td>
<td>0.001***</td>
<td>0.506</td>
<td>0.000***</td>
<td>0.305</td>
<td>0.001***</td>
<td>0.177</td>
<td>0.055</td>
</tr>
<tr>
<td>PM</td>
<td>0.553</td>
<td>0.000***</td>
<td>0.823</td>
<td>0.000***</td>
<td>0.542</td>
<td>0.000***</td>
<td>0.449</td>
<td>0.000***</td>
</tr>
<tr>
<td>Year (ref:2013)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year (2014)</td>
<td>0.162</td>
<td>0.077</td>
<td>0.154</td>
<td>0.086</td>
<td>0.147</td>
<td>0.140</td>
<td>0.125</td>
<td>0.162</td>
</tr>
<tr>
<td>Year (2015)</td>
<td>0.046*</td>
<td>0.619</td>
<td>0.001</td>
<td>0.989</td>
<td>0.141</td>
<td>0.152</td>
<td>0.007</td>
<td>0.938</td>
</tr>
<tr>
<td>Region (ref:east)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>North</td>
<td>0.074</td>
<td>0.468</td>
<td>0.083</td>
<td>0.409</td>
<td>0.083</td>
<td>0.419</td>
<td>-0.099</td>
<td>0.387</td>
</tr>
<tr>
<td>South</td>
<td>0.318</td>
<td>0.002</td>
<td>0.361</td>
<td>0.000***</td>
<td>0.320</td>
<td>0.002**</td>
<td>0.149</td>
<td>0.194</td>
</tr>
<tr>
<td>West</td>
<td>0.731</td>
<td>0.000</td>
<td>0.695</td>
<td>0.000***</td>
<td>0.742</td>
<td>0.000***</td>
<td>0.927</td>
<td>0.000***</td>
</tr>
<tr>
<td>Interactions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strava*am</td>
<td>-0.181</td>
<td>0.000***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strava*pm</td>
<td>-0.200</td>
<td>0.000***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strava*2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strava*2015</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strava*North</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strava*South</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strava*West</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept (con)</td>
<td>3.057</td>
<td>0.000</td>
<td>2.882</td>
<td>0.000***</td>
<td>3.028</td>
<td>0.000</td>
<td>3.099</td>
<td>0.000***</td>
</tr>
<tr>
<td>Dispersion</td>
<td>1.074</td>
<td>1.120</td>
<td>1.081</td>
<td>1.081</td>
<td>1.141</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

- Strava shows good correlation with observed cycle counts
- The correlation is higher the more we aggregate the observations
- These correlations change depending on different factors
- This seems to correspond with what has been found in the literature
Thank you for your attention.

The data used are available from the Urban Big Data Centre.