The Statistics of Cycling

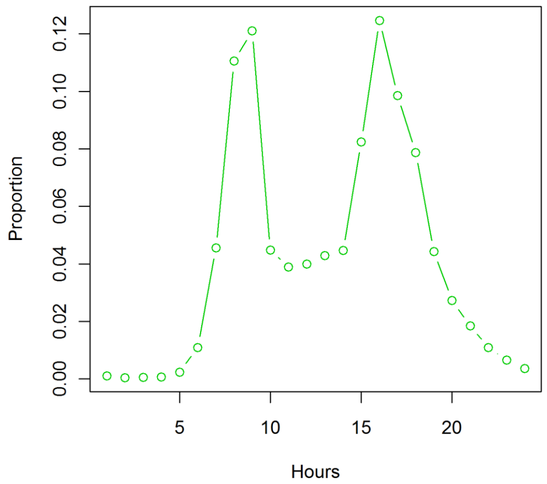
Matthew Arnold
Grey College

The data is supplied by Sustrans, a charity that promotes sustainable transport in the UK. Sustrans is responsible for the planning and implementation of cycle routes (including the National Cycle Network) in the UK. Located along these routes are counters that record the number of bicycles that cross the loop of wire buried underneath the path.

The photographs show the counter and box, and highlight the induction loop in the ground. Data files for each counter contain the number of counts in 2 directions per hour. Data is usually available for 2004-mid 2009; for some counters data is available back to 1999.

1. Usage profiles

At each counter we create an average day profile, showing the proportion of the daily count at that hour. We can then plot these as a daily profile graph. We also separate these daily profiles into weekday and weekend profiles, where we may see different usage patterns. Shown below is the daily usage profile for a counter in Shropshire on weekdays.



We similarly construct a yearly profile for a counter by considering the average proportion of usage per month.

2. Clustering and classification

We compare the daily profile graphs for two counters using the Euclidean distance metric.

$$D_{ij} = \left(\sum_{k}^{24} (x_{ik} - x_{jk})^2\right)^{\frac{1}{2}} = \| x_i - x_j \|$$

We find clusters of similar daily profiles by applying k-means clustering; the algorithm executes in two repeated steps after selecting $m_1, \ldots, m_k$ initial starting means

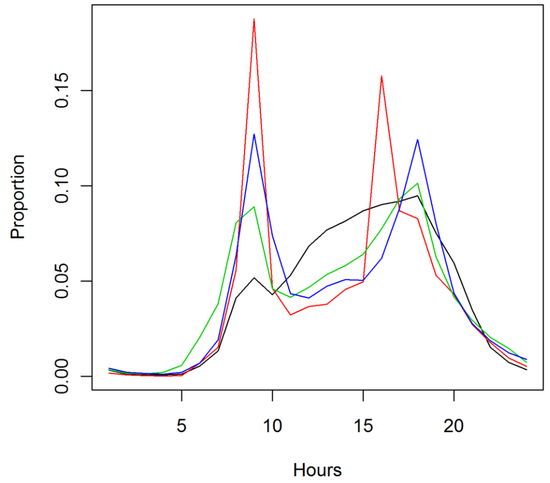
Assignment: Assign each point to its closest mean according to the Euclidean metric.

$$S_i^{(t)} = \left\{ x_j : \| x_j - m_i^{(t)} \| \le \| x_j - m_{i^*}^{(t)} \| \; \forall i^* = 1, \ldots, k \right\}$$

Update: Update the means m_i to be the centroid of the observations in the cluster.

$$m_i^{(t)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

The algorithm ends when the assignment does not change upon further iterations.
Our clustering shows 4 different classifications of counters. By inspection and our intuition, the black graph represents the average day at a leisure route counter; the red near a school, blue predominantly commuters and green a hybrid of leisure and commuter activity.



3. Linking to explanatory variables

We perform a Fisher's exact test on a 4x2 contingency table of our counter classifications against 30 binary variables. We calculate the probability of observing such a set of values using the hypergeometric distribution under the null hypothesis that classification is independent of the state of an explanatory variable.As an example we create a contingency table and test to see if a route being "trafficfree" affects the counter classification.

Observed counts	R	G	B	B
Traffic free	4	27	13	29
Not traffic free	1	16	14	3

giving a p value of 0.002, suggesting that there may be some relationship between traffic free state and a counter's classification.

We also fit a binomial logit model (BLM) to categorise a counter given knowledge of the explanatory variables at that location. BLMs are generalised linear models (GLM). Each response category is compared to a baseline category J (often either the last or most common one). Letting

$$\pi_j(\mathbf{x}) = P(Y = j \mid \mathbf{x})$$

for fixed explanatory variables $\mathbf{x}$, the model

$$\log \frac{\pi_j(\mathbf{x})}{\pi_J(\mathbf{x})} = \alpha_j - \boldsymbol{\beta}_j^T \mathbf{x} \qquad (1)$$

simultaneously describes the effect of $\mathbf{x}$ on J-1 logits, which vary according to the response paired with the baseline. $\boldsymbol{\beta}$ is determined via maximum likelihood estimation. We find the response probabilities given $\mathbf{x}$ using

$$\pi_j(\mathbf{x}) = \frac{\exp(\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x})}{1 + \sum_{k=1}^{J-1} \exp(\alpha_k + \boldsymbol{\beta}_k^T \mathbf{x})}$$

with $\alpha_J = 0$ and $\boldsymbol{\beta}_J = \mathbf{0}$ which follows from (1) and we also that

$$\sum_j \pi_j = 1.$$

Further work will analyse the response probabilities and the significance of the model.

"If you worried about falling off the bike, you'd never get on."
Lance Armstrong

sustrans JOIN THE MOVEMENT

Durham University