Zipf's law in activity schedules

Extended abstract submitted to hEART 2016

Wim Ectors*, Bruno Kochan, Davy Janssens, Tom Bellemans & Geert Wets
Transportation Research Institute, Hasselt University, Wetenschapspark 5 Box 6, B-3590 Diepenbeek, Belgium

March 15, 2016

1 Background

The transportation research community has been working on traffic demand models for many years. Modeling people's behavior is an extremely complex, multidimensional process. The number of degrees of freedom in a person's activity schedule is enormous. However, as we will demonstrate, the frequency of occurrence of day-long activity schedules obeys a remarkably simple, scale-free distribution.

This distribution, commonly referred to as Zipf's law, obeys a power-law which has been observed in many natural and social processes. It was actually first observed by Felix Auerbach in 1913 (Auerbach, cited in Zipf 1949, Newman 2005). He discovered that city size is governed by such a power-law. Willis (Willis, cited in Chen 1980) noted in 1922 that the size distribution of biological genera follows a power-law distribution. Zipf, an American linguist, described a power-law distribution in word frequency in 1949 (although it had first been noticed by Estroup in 1916 (Estroup, cited in Ki Baek et al. 2011)). Zipf famously investigated this distribution more in detail, revealing that the same power-law distribution holds for a large number of events in different domains, ranging from sizes of earthquakes, annual income of companies, solar flares, to the number of citations received on papers (Fujiwara, 2004; Furusawa and Kaneko, 2003; Maillart et al., 2008; Newman, 2005; Okuyama et al., 1999).

The rank-size interpretation of Zipf's law is most common. For example: within the context of city sizes, the size of a city at rank i varies as $1/i$. The second largest city is then half the first city's size, the third largest one-third its size etc.:

$$f(r_i) = \frac{f(r_1)}{r_i} \tag{1}$$

where f represents *frequency* and r the *rank*. In other words, the size of a city is inversely proportional to its rank. The data obeys a power-law distribution with exponent close to 1.0.

Mentions of Zipf's law within the domain of transportation sciences are very thin. Power-law-like distributions have been evidenced in displacement distance, gyration radius and location visiting frequency (González et al., 2008), as well as in location visiting duration (Brockmann et al., 2006). Yang et al. (2014) found power-law distributions in bus transport networks. Guidotti et al. (2015); Klafter et al. (1996); Song et al. (2010) also provide noteworthy contributions within this topic.

2 Aims

The aims of this research are:

- to provide evidence for the generalization of a rank-frequency Zipf's law in activity schedule frequencies
- to test the law's dependency on the aggregation level of activity types

*Corresponding author. E-mail address: wim.ectors@uhasselt.be Tel.: +32-11-269114

- to test the law's validity for each day of the week

3 Methodology

The methods employed to fulfill the aims are:

- **Visually observing activity schedule frequency distributions for different study area's:**

  Activity schedules were generated for the OVG 3.0-4.5 (Flanders, Belgium), OViN 2013 (Netherlands) and NHTS 2009 (U.S.) travel survey datasets. Frequency tables were generated and normalized (schedule with the highest frequency as the denominator). Figure 1 illustrates the result.

- **Statistically determining power-law fits and calculating parameter uncertainties using the R package "poweRlaw". Additionally, the power-law parameters were re-estimated on subsets of the dataset (according to day of the week) and used to evaluate the validity of Zipf's law:**

  In order to claim with relative certainty that the empirical Zipf's law can be observed in the domain of transportation behavior, a power-law distribution should be fitted to the data. Usually this is performed as a linear regression (least-squares) with transformed variables, however this method is flawed as explained by Clauset et al. (2009); Newman (2005); Urzúa (2011). The slope estimate may suffer from systematic, large errors. Clauset et al. (2009) proposed a method based on maximum-likelihood fitting methods using the Kolmogorov-Smirnov goodness-of-fit statistic. The R package called "poweRlaw" (Gillespie, 2015) was developed to automate this method.

  The famous exponent $\alpha' = 1$ of Zipf's law refers to the *cumulative* distribution function, which relates as $\alpha' = \alpha - 1$ where α is the exponent in the probability distribution function $p(x) = Cx^{-\alpha}$. Therefore, as will be observed in the results, an $\alpha = 2$ would confirm Zipf's law in the data. Table 1 presents results of this experiment. Figure 2 illustrates a power-law fit.

- **Using different sets of activity encoding and testing the validity of Zipf's law depending on the used encoding set**

  Different activity encoding aggregation levels were created. Starting from the original activity type encoding in the NHTS 2009 dataset, four more sets of encoding were proposed, each aggregating some activity types or grouping them somewhat differently. This approach corresponds to constructing an encoding tree and pruning the branches to increase the aggregation level. Figure 3 illustrates the result.

4 Results and conclusion

Figure 1 and Figure 2 provide evidence for the generalization of a rank-frequency Zipf's law in activity schedule frequencies. A strong support for this claim is given in Table 1, yielding satisfying exponent estimates and statistical evidence in the form of a weak prove based on hypothesis tests. Figure 3 presents the result of testing the law's dependency on the aggregation level of activity types. One observes that only in case of the most severe aggregation in *Code_cat* Zipf's law breaks down quite soon; for the other cases its validity is illustrated for the majority of observations. Therefore, it is concluded that Zipf's law does not strongly depend on the activity encoding. Finally, Table 1 also lists exponent estimates that confirm a good fit of Zipf's law across the days of the week.
Further research will continue to build evidence for scale-free distributions in the domain of transportation. It will attempt to find the limits of Zipf's law in more disaggregated data, as well as suggesting a cause for the remarkable simple distribution.

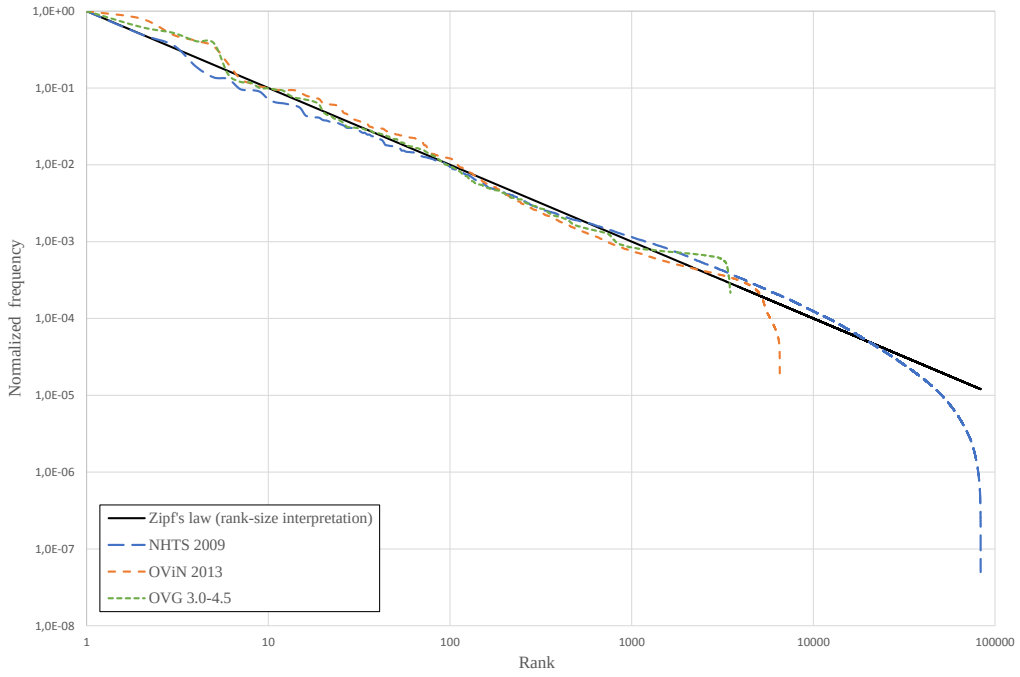

Figure 1: Investigating Zipf's law in day-long activity schedules, for different study areas. The horizontal axis shows the rank (in a list of descending frequency), while the vertical axis shows the normalized frequency of activity schedules.

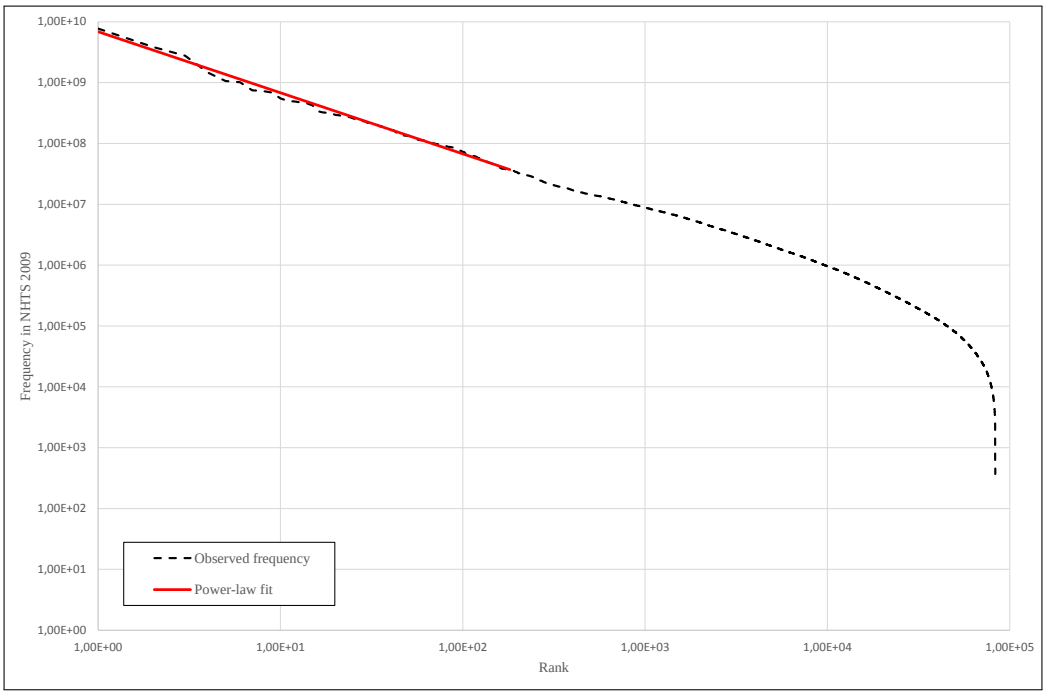

Figure 2: Power-law fit for the activity schedules in NHTS 2009, according to the poweRlaw R package. In this rank (cumulative) plot, the fit is $Cr^{-2.003+1}$ with r the rank, yielding an exponent very close to Zipf's value of 1. The optimization of Xmin using the Kolmogorov-Smirnov goodness-of-fit statistic is responsible for the limited range of the power-law fit. However, as one may observe, this power-law trend may extend to a much larger range, although perhaps at a minor cost of accuracy of the fit.

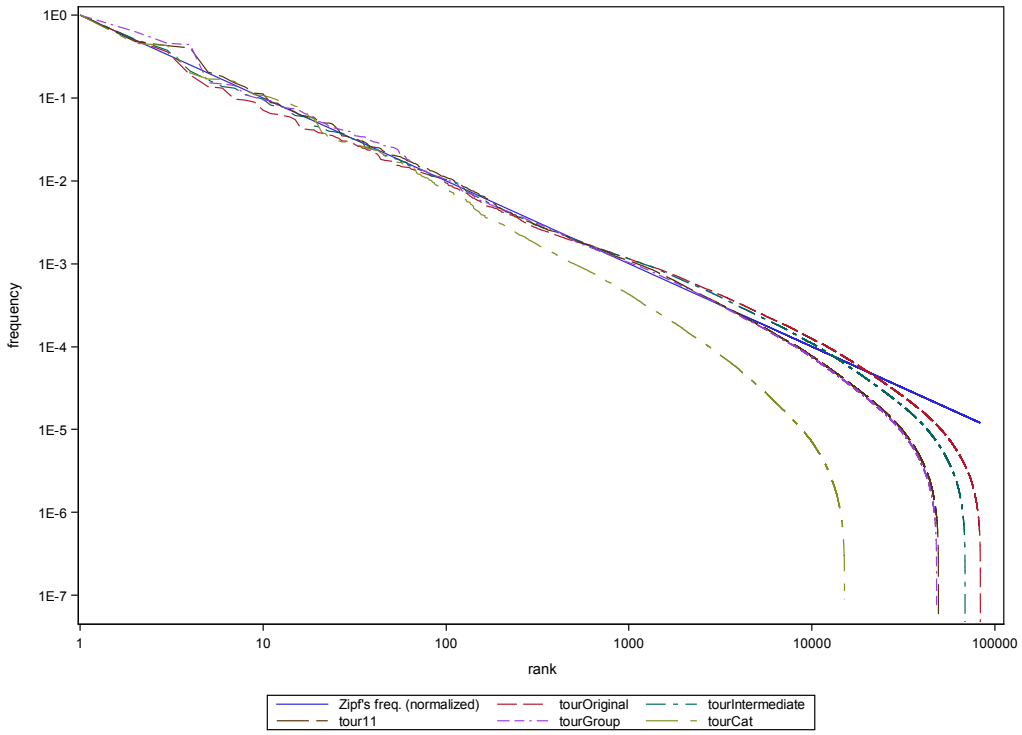

Figure 3: Investigating Zipf's law in day-long activity patterns from the NHTS 2009 dataset, with different activity-type aggregation levels. The horizontal axis shows the rank (in a list of descending frequency), while the vertical axis shows the normalized frequency. The encoding schemes tourOriginal, tourIntermediate, tour11, tourGroup and tourCat have respectively 37, 18, 10, 10 and 3 distinct activity type-alternatives.

Table 1: Results of the PoweRlaw package for different (sub-)datasets. The bootstrapping procedure samples (with replacement) from the dataset and re-infers the parameters, providing an indication of the parameter uncertainty. *AM* is the arithmetic mean and *SD* the standard deviation of 5000 bootstrapping simulations. The P-value refers to a hypothesis test with H_0: a power-law cannot be ruled out, and H_1: a power-law is ruled out. We reject H_0 if $p < 0.10$.

		PoweRlaw parameter estimation				Bootstrapping uncertainty evaluation		
Dataset	Subset	α	Xmin	Cum. Pct discarded	n_{tail}	AM(α)	SD(α)	P-value
NHTS 2009	all	2.003	36809977	55%	181	2.006	0.070	0.255
OViN 2013	all	1.885	1325	20%	721	1.877	0.058	0.005
OViN 2013 (OVG encoding)	all	1.830	2378	22%	406	1.862	0.052	0.336
OVG 3.0-4.5	all	1.947	2	23%	529	1.953	0.051	0.149
NHTS 2009	Monday	2.290	46616705	67%	22	2.270	0.359	0.831
NHTS 2009	Tuesday	2.161	35581917	67%	26	2.182	0.236	0.820
NHTS 2009	Wednesday	2.152	45646004	68%	20	2.172	0.267	0.679
NHTS 2009	Thursday	2.088	48120314	71%	17	2.140	0.282	0.221
NHTS 2009	Friday	2.279	34509610	72%	28	2.284	0.250	0.901
NHTS 2009	Saturday	2.182	61045896	76%	15	2.176	0.288	0.134
NHTS 2009	Sunday	2.091	52160661	66%	21	2.060	0.200	0.982

5 Bibliography

Auerbach, F. (1913), 'Das Gesetz der Bevölkerungskonzentration', *Petermanns Geographische Mitteilungen* **59**, 74–76.

Brockmann, D., Hufnagel, L. and Geisel, T. (2006), 'The scaling laws of human travel.', *Nature* **439**(7075), 462–5.
URL: *http://dx.doi.org/10.1038/nature04292*

Chen, W.-C. (1980), 'On the Weak Form of Zipf's Law', *Journal of Applied Probability* **17**(3), 611–622.

Clauset, A., Shalizi, C. R. and Newman, M. E. J. (2009), 'Power-Law Distributions in Empirical Data', *SIAM Review* **51**(4), 661.
URL: *http://link.aip.org/link/SIREAD/v51/i4/p661/s1&Agg=doi*

Estroup, J. (1916), *Les Gammes Sténographiques*, 4th editio edn, Institut Stenographique de France, Paris.

Fujiwara, Y. (2004), 'Zipf law in firms bankruptcy', *Physica A: Statistical Mechanics and its Applications* **337**(1-2), 219–230.
URL: *http://www.sciencedirect.com/science/article/pii/S0378437104001165*

Furusawa, C. and Kaneko, K. (2003), 'Zipf's Law in Gene Expression', *Physical Review Letters* **90**(8), 1–11.

Gillespie, C. S. (2015), 'Fitting Heavy Tailed Distributions: The poweRlaw Package', *Journal of Statistical Software* **64**(2), 1–16.
URL: *http://www.jstatsoft.org/v64/i02*

González, M. C., Hidalgo, C. A. and Barabási, A.-L. (2008), 'Understanding individual human mobility patterns', *Nature* **453**(7196), 779–782.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/18528393*

Guidotti, R., Trasarti, R. and Nanni, M. (2015), 'TOSCA : TwO-Steps Clustering Algorithm for Personal Locations Detection', *Proceedings of the 23nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (January).

Ki Baek, S., Bernhardsson, S. and Minnhagen, P. (2011), 'Zipf's law unzipped', *New Journal of Physics* **13**(4), 043004.
URL: *http://stacks.iop.org/1367-2630/13/i=4/a=043004*

Klafter, J., Shlesinger, M. F. and Zumofen, G. (1996), 'Beyond Brownian Motion', *Physics Today* **49**(2), 33.
URL: *http://pmc.polytechnique.fr/pagesperso/dg/cours/biblio/Physics Today, February 33 (1996) Klafter, Shlesinger, Zumofen [Beyond Brownian Motion].pdf\nhttp://link.aip.org/link/PHTOAD/v49/i2/p33/s1&Agg=doi*

Maillart, T., Sornette, D., Spaeth, S. and von Krogh, G. (2008), 'Empirical Tests of Zipf's Law Mechanism in Open Source Linux Distribution', *Physical Review Letters* **101**(21), 218701.
URL: *http://link.aps.org/doi/10.1103/PhysRevLett.101.218701*

Newman, M. (2005), 'Power Laws, Pareto Distributions and Zipf's Law', *Contemporary physics* **46**, 323–351.
URL: *http://www.tandfonline.com/doi/abs/10.1080/00107510500052444*

Okuyama, K., Takayasu, M. and Takayasu, H. (1999), 'Zipf's law in income distribution of companies', *Physica A: Statistical Mechanics and its Applications* **269**(1), 125–131.

Song, C., Koren, T., Wang, P. and Barabasi, A.-L. (2010), 'Modelling the scaling properties of human mobility', *Nature Physics* **6**(10), 1–6.
URL: *http://dx.doi.org/10.1038/nphys1760\npapers2://publication/doi/10.1038/nphys1760*

Urzúa, C. M. (2011), 'Testing for Zipf's law: A common pitfall', *Economics Letters* **112**(3), 254–255.
URL: *http://dx.doi.org/10.1016/j.econlet.2011.05.049*

Willis, J. C. (1922), *Age and Area*, Cambridge University Press.

Yang, X. H., Chen, G., Chen, S. Y., Wang, W. L. and Wang, L. (2014), 'Study on some bus transport networks in China with considering spatial characteristics', *Transportation Research Part A: Policy and Practice* **69**, 1–10.
URL: *http://dx.doi.org/10.1016/j.tra.2014.08.004*

Zipf, G. K. (1949), *Human Behaviour and the Principle of Least Effort*, Addison-Wesley, Reading.