Measuring Attraction and Redistribution of Institution-Based Movements

M. Pistner[1] and C. Andris*[1]

[1]The Pennsylvania State University
Email: {map5672, *clio}@psu.edu

Abstract

How can we uncover the human dynamics behind moving to and from a large rural university? We apply clustering and socio-economic modelling statistical methods to measure changes in the Pennsylvania State University's population draw and redistribution power from 1995-2015, focusing on the U.S. Mid-Atlantic Region. Our data contain aggregate hometowns of 225,000 students and aggregate locations of 295,000 alumni at the U.S. zip code and country level. Our results show new alumni bases, the persistent role of student-rich suburbs, and the geographic fanning of student migrants.

Keywords: Institutions, Circular Statistics, Flow Data, Migration.

1. Motivation

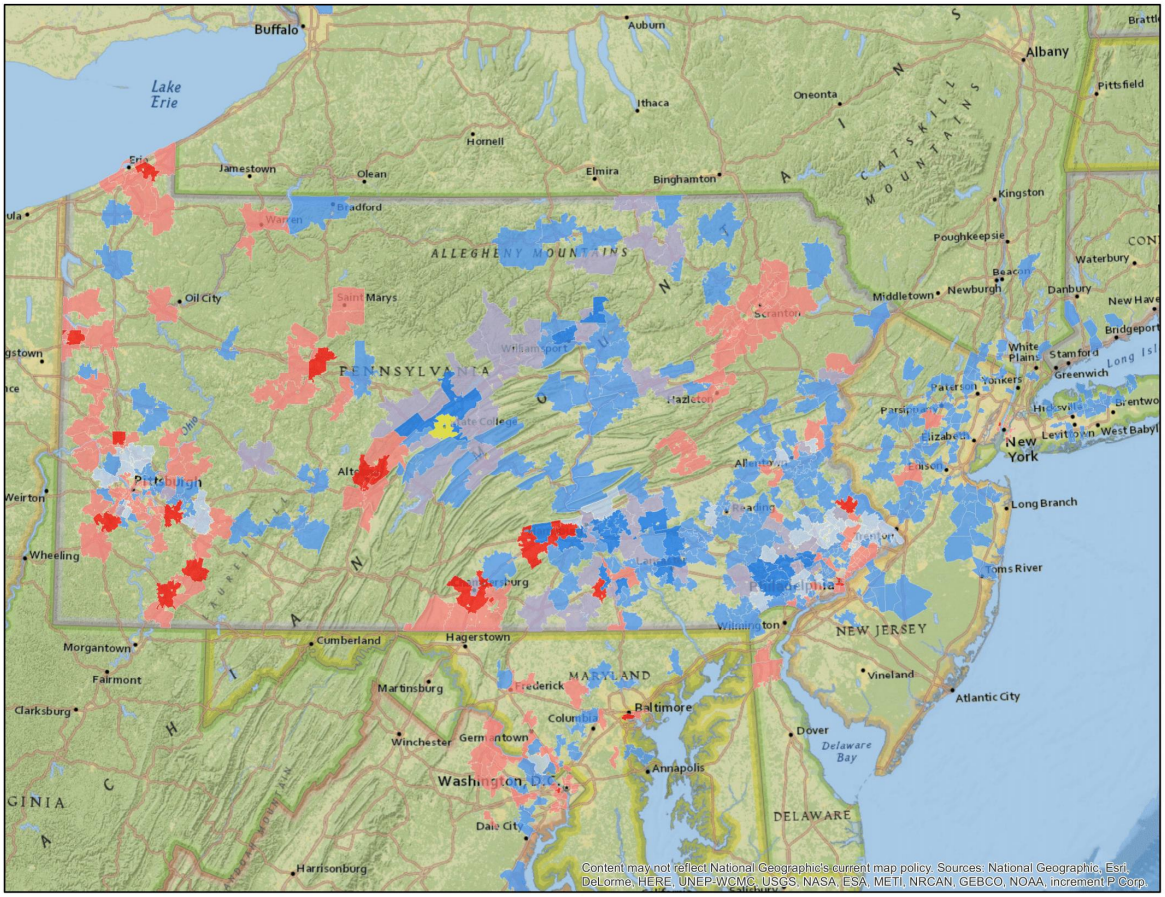
Higher education institutions (colleges and universities) are a special type of spatial feature that attract students from different locales, and after a certain period of time and skill-attainment, redistribute these students to (presumably) a new set of locales. These schools have a unique pull and push power: they attract students from a set of geographies which act as a "potential" energy for the university's impact on people from those places. After the student attends the university, this potential energy becomes "kinetic" energy, as the graduate takes his or her experience to another place (or back home), thus enriching their cities, towns and neighbourhoods with skills and, to some extent, the school's brand.

What does this distribution process look like? We use data from the Pennsylvania State University (Penn State) that contain hometowns of undergraduate and graduate applicants categorized by year applied for 1995-2015 as well as alumni home location at the U.S. zip code level categorized by graduation year for 1995-2015. In 1995, approximately 27,000 undergraduate and 12,000 graduate students applied and 11,000 alumni registered as having graduated in 1995. 2015 saw 57,000 undergraduate and 16,000 graduate applicants, and 17,000 alumni for the class year 2015. We apply K-means clustering, socio-economic modelling using Poisson regression, and circular statistics via chi-squared tests to this flow dataset in order to gage changes in an institution's draw power and redistribution power over time. We find that students hail from suburban regions and upon graduating, tend to move to both urban and rural locations.

2. Geographical Displacement by Cohort (Year)

When a place sends students to an institution, will it see the same number of students return to the place? We created a temporal signature vector where each element represents the difference of number of applicants to be in *the class of* (for example) 2000 (i.e. applied around 1995) vs. alumni that graduated in 2000 (their cohort's sets of post-graduate locations). A vector is created for each U.S. zip code, and has 21 distinct elements, one corresponding to each year. We used k-means clustering to classify these signature vectors. A total of nine clusters were selected based off diagnostic results (Table 1). Given the four-year time lag, we can only use applicant data until year 2011 (for the class of 2015, for which we have alumni).

Two large clusters were characterized by little or no net fluctuations. Three clusters showed a strong student applicant pool, that was not replenished in alumni four years later. Students are typically drawn from the suburbs, but alumni are more likely to return to the city centre. Interestingly, Pennsylvania cities known for their mixed economic downturn and resurgence behaviours, Pittsburgh, Scranton, and Erie (Laboon, 2014; High, 2015) are characterized by a large alumnus draw relative to the students they send. Conversely, Philadelphia and New Jersey are typically consistent student bases that send many students each year, but receive proportionally fewer alumni in return—perhaps indicating that graduates can find jobs in new locations.



Legend

- State College
- Large Alumni Base (consistent)
- Large Alumni Base (variable)
- Student-to-Alumni Shift (variable)
- Student Base (varable)
- Large Student Base (variable)
- Student Base (consistent)

Figure 1: Maps of zip codes clustered by net temporal flow shows alumni more prominently in the western part of the state.

3. Socio-Economic Model

The effects of Euclidean distance at the hometown (zip code level) and national levels were tested as drivers of undergraduate enrolees, graduate enrolees, and alumni movements. For U.S. locations, income was measured as the median income per zip code as reported by Population Studies Center

(PSC, 2017). For international locations, income was approximated by national gross domestic product (GDP) per capita.

Cluster	Count of Entities	Mean	Standard Deviation	Assigned Typology
1	346	1.60	1.76	Student-to-Alumni Shift
2	229	-4.66	2.26	Large Alumni Base (consistent)
3	2	473.59	163.80	State College
4	16711	-0.07	0.37	Unused (*low participation*)
5	26	38.66	14.76	Large Student Base (variable)
6	35	2.76	5.02	Student Base (variable)
7	54	11.80	4.65	Student Base (consistent)
8	22	-18.94	7.13	Large Alumni Base (variable)
9	21326	0	0	Unused

Table 1. Statistics by cluster

Using Getis-Ord Hot spot detection, we found three pools of concentrated interaction with the university: large in-state metropolises Pittsburgh (PGH) and Philadelphia (PHL) as well as Centre County (CTR), the school's locale and tagged these as indicator (dummy) variables, producing the model:

$$movement_i = \beta_0 + \beta_1 \log(d)_i + \beta_2 \log(inc)_i + \beta_3 I_{CTR_i} + \beta_4 I_{PGH_i} + \beta_5 I_{PHL_i} + \varepsilon_i \qquad \text{Equation 1}$$

Where d is distance, inc is income, β is a coefficient, ε_i is an error term, and I denotes an indicator (dummy) variable that can take on a value [0,1]. We fit this model separately for in-state, out-of-state, and international movements. Ordinary least squares (OLS), hurdle (Zeileis et al. 2016) and zero-inflated regression models (Rideout et al. 1998) were tested, and each produced higher root mean square error (RMSE) values than a Poisson regression model (Table 2). Poisson regression simplified model assumptions when compared to the zero-inflated and hurdle models.

Method	In-state	Out-of-state	International
OLS	6.26	1.04	10.94
Poisson	4.26	0.93	5.64
Hurdle	4.30	1.03	8.64
Zero-inflated	4.68	0.93	8.41

Table 2. Mean RMSE for undergraduate models

The effect of distance on in-state and out-of-state undergraduates is consistent over time: with increased distance, there are fewer applicants. For undergraduate students, the mean distance coefficients are -0.342 (s.d. 0.078) for in-state students, -0.813 (s.d. 0.065) for out of state students and 0.991 (s.d. 0.38) for international students. Distance decay is only an issue once state boundaries are crossed. International undergraduates are less hampered by distance over time, which perhaps indicates that if one commits to traveling internationally for college, the actual distance is not a factor. Distance poses less of a barrier for out-of-state graduate students, especially in recent years. International graduate students are becoming more sensitive to distance as a product of time. Positive coefficients are driven by a rise in applications originating from Asia.

Higher income areas are producing an increasing amount of in-state and out-of-state undergraduate students, potentially due to increasing tuition costs and Penn State's status as one of the most expensive public universities nationwide (Friedman 2016). For international undergraduates, there is a small, but consistent, positive correlation between income and number of enrolees. In contrast, a nation's income

does not appear to effect graduate student production, perhaps as most graduate programs offer tuition remission for their students. Wealthier locations in the United States consistently attract more alumni than their less-wealthy counterparts, which may be due to young alumni moving to downtown city areas.

4. Circular Statistics

To study the radial distribution of applicants, we created eight bins of equal angular ranges and populated the bins by the sum of movements within the angle ranges. We then used a circular chi-squared test (X^2) to test the degree to which these movements favoured a certain angle. (We found our statistics to be invariant to the start angle of the bins and the total number of overall bins). Then, the circular X^2 test statistic (Eq. 2) can be calculated accordingly:

$$X^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i} \xrightarrow{d} X^2_{\#bins-1} \qquad \text{Equation 2}$$

where o_i is the observed count in bin *i* and e_i is the expected count in bin *i*. Inference on this test statistic was based on the $X^2_{\#bins-1}$ distribution. Total enrolees by bin was calculated for overall enrolment (Table 3) and the hypothesis of uniformity is rejected for every year at $\alpha = 0.05$ even considering multiple comparisons. When controlling for sample size (by paring samples down to the global minimum for any type-year: 7500 for alumni, 6800 for undergrads, and 2500 for graduate students), we find that non-uniformity is most pronounced for alumni and least pronounced for undergraduates (Figure 2), but that no group fluctuates annually in their degree of angular uniformity. However, graduate students seem to be becoming less uniform most rapidly over time (Figure 2).

An eastern proclivity is the source of non-uniformity: including Philadelphia and New Jersey, as well the international student powerhouses of India and China. Graduate student origins were highly clustered in 1995, but now exhibit a wider variety of regions especially from overseas. In 1995, the majority (1025 students) attended graduate school from the University's zip code. In 2015, only 115 attended from that zip code (Figure 3).

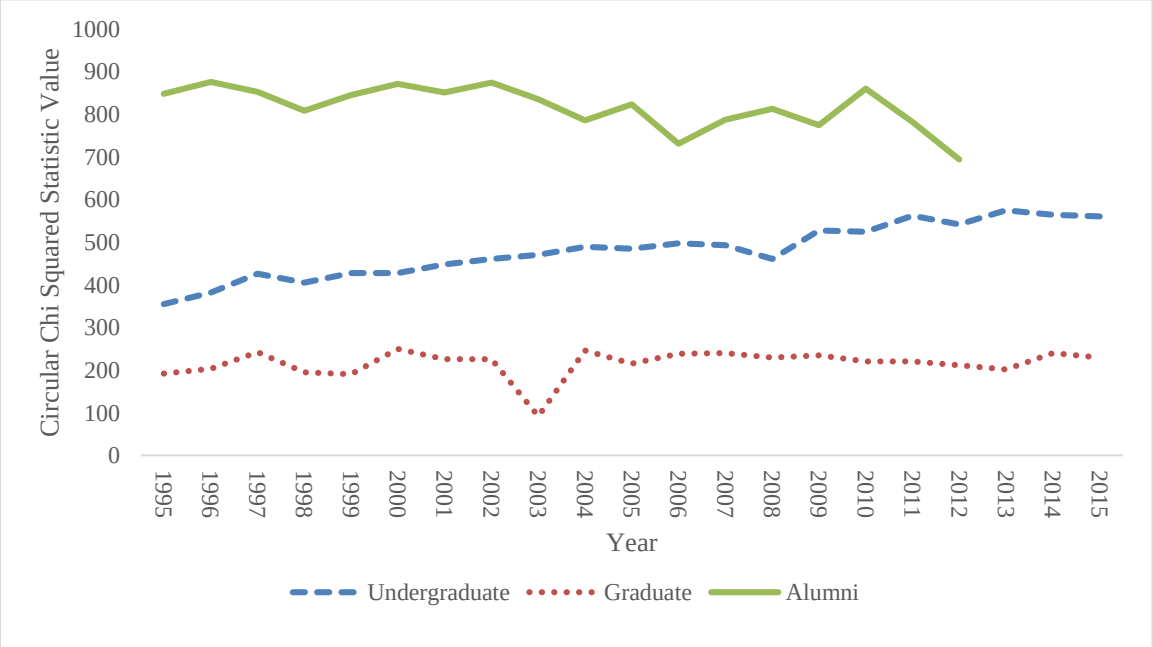


Figure 2: Circular X^2 statistics with controlled sample sizes show relative stability over time.

Bin	Undergraduate enrolees	Graduate enrolees	Alumni
(337.5, 22.5]	2516	710	1258
(22.5, 67.5]	5696	4257	5831
(67.5, 112.5]	76312	31497	116857
(112.5, 157.5]	24612	11221	38799
(157.5, 202.5]	8273	4168	19354
(202.5, 247.5]	4318	3483	15961
(247.5, 292.5]	30296	15251	55302
(292.5, 337.5]	3607	1269	8049

Table 3. Magnitudes of entities in different angular bins

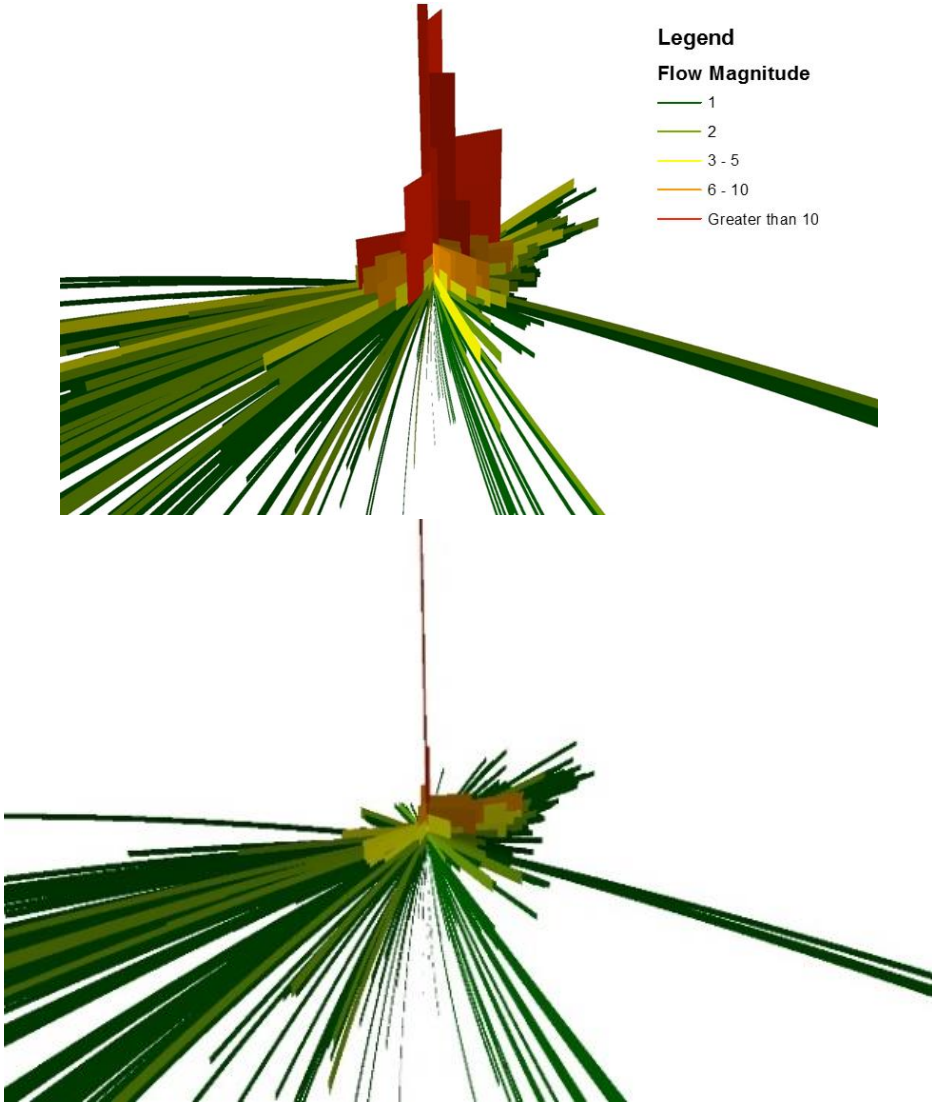


Figure 3: Flow magnitude graphs for graduate enrolees in 1995 (top) and 2015 (bottom) show a de-clustering of students from near the university.

5. Conclusions

In summary, we used a spatial dataset of over 500,000 institution-based migrant flows and found that certain local urban and suburban areas attract students but do not replenish the population with alumni. We found that within state borders and beyond international borders, distance does not play a role in attracting undergrads, but that only in out-of-state U.S. does the distance matter. Finally, we discovered that alumni have the most variegated radial distribution, followed by graduate students, and then undergraduates. Although these angular distributions stay relatively constant over time, graduate student enrolees especially hail from a wider and more dispersed set of geographies than in the past—more than any other group.

6. Acknowledgements

This work was supported by the Pennsylvania State University and the National Science Foundation under an IGERT award #DGE-1144860, Big Data Social Science. This project was funded by the Wilson Initiation Grant from the College of Earth and Mineral Sciences at Penn State. Data is graciously provided by Penn State Admissions Office and Penn State Alumni Association.

7. References

Friedman, J. 2016. 10 Colleges with the Highest Tuition for In-State Students. *U.S. News*. Accessed online Jun. 12, 2017. Available from: https://www.usnews.com/education/best-colleges/the-short-list-college/articles/2016-05-03/10-colleges-with-the-highest-tuition-for-in-state-students

High, S. 2015. *Industrial sunset: the making of North America's rust belt, 1969-1984*. Toronto: University of Toronto Press.

Laboon, P. 2014. What Entrepreneurs Can Learn from Pittsburgh's Renaissance. *Forbes Magazine*. Accessed online Mar. 1, 2017. Available from: https://www.forbes.com/sites/theyec/2014/09/08/what-entrepreneurs-can-learn-from-pittsburghs-renaissance/#31227d98fc2c

Population Studies Center (PSC), University of Michigan, 2017. Zip Code Characteristics: Mean and Median Household Income. Accessed online Mar. 1, 2017. Available from: http://www.psc.isr.umich.edu/dis/census/Features/tract2zip/.

Ridout, M., Demétrio, C.G. and Hinde, J., 1998. Models for count data with many zeros. In: *Proceedings of the XIXth International Biometric Conference*, **19**, pp. 179-192.

Zeileis, A., Kleiber, C. and Jackman, S., 2008. Regression models for count data in R. *Journal of Statistical Software*, **27**(8), pp.1-25.