Good start.

Your discussion of sample size seems rather confusing to me.

Also don't forget to do some analyses treating your registrar data as a (post-) stratiied sample from the full target population of off-campus students.

the GPS stuff is very interesting

I look forward to an interesting and informative report.

-BJ

36-303 Team G

Spatial and Analytical Study of Student Housing at Carnegie Mellon University

A DRAFTY DRAFT

Ariel Liu, Sam Lavery, Alejandra Munoz, Terra Mack, Shannon Lauricella

4/6/2012

Introduction

In the study, “A Quasi-Experimental Approach to Estimating the Impact of Collegiate Housing,” Ryan Yeung observed that students “from the residence halls to off-campus housing…become less integrated into the academic and social systems of the college.”[1] Research sponsored by the U.S. Department of Education upholds this idea that living on campus provides a stronger support system, more engagement in educational practices, and increased social interaction.[2] These key factors provide “the single most consistent within-college determinant of the impact of college.”[3]

However, the numbers do not seem to match this sentiment. According to figures published in the Digest of Education Statistics, 86.2% of undergraduate students in the United States live off-campus, with 55.2% not living with their parents[4], suggesting that students and colleges do not necessarily “subscribe to this theory of the benefits of on-campus housing.”[5] Recent articles published by *The Tartan*, Carnegie Mellon University’s student-run newspaper, indicate that more and more students at CMU are actively looking to move off-campus.[6]

This discrepancy between the reported increased well-being of students living on-campus versus the number of students living off-campus served as a motivating factor for our survey. Do students care about integrating the supposed benefits of living on-campus when choosing an off-campus residence? Considering factors like neighborhood, proximity to peer groups, and relative distance to campus and campus amenities, like shuttle stops may increase academic well-being and foster social interaction, perks that are often attributed to living on-campus.

Our survey seeks to answer questions about the dynamics of student housing at CMU. We are particularly interested in investigating the correlation between where students choose to live and what they choose to study. The results of the survey will be valuable for students in finding

[1] (Yeung, 2011)
[2] (Schudde, 2011)
[3] (Pascarella & Terenzini, 2005)
[4] (Snyder, Dillow, & Hoffman, 2009)
[5] (Yeung, 2011)
[6] (Fitzgerald, 2006)

move these into the text, replacing the footnotes there (to follow APA style) neighborhoods within the city that are popular with students like themselves as well the university in planning shuttle routes, campus police coverage, and future housing projects

Here, we will include a quick summary of the main results from Section 4 (which we have not finished yet).

Methods

In order to explore our question, we needed to understand the population to sample, what questions to pose to the population, and how to process the data.

Target Population and Sampling Frame

The target population in our study is undergraduate and graduate students enrolled at CMU that live outside of the main campus Pittsburgh. It is the same population that we are looking to make inferences about from our survey. The target population differs from the sampling frame in that addresses are self-reported to the registrar and students may neglect to update their address. Thus, we will not have access to information for the entire population enrolled, so our sampling frame will include only those students who comply with the registrar's office or volunteered their information to CMU.

ok

Sample Size

you are not telling me why you are calculating sample size, or what survey design you are calculating sample size for.

According to the CMU Factbook, there are 2,252 undergraduates living off-campus and 5,769 graduates living off-campus.[7]

To calculate the sample size, we selected the following question: 'Is this person a member of CIT (Carnegie Institute of Technology)?'. Then, from the Factbook, we used the head count of students in each college enrolled in Fall 2011 in Pittsburgh to calculate the proportion of the student body represented by students enrolled in CIT (p).

[7] (Office of Institutional Research and Analysis, 2012)

The total head count of students: 10,957

The head count for CIT students: 3,217

The proportion of CIT students out of total students: $p = \frac{3,217}{10,957} = .293 = 29.3\%$

Total population size of students living off-campus: $2,252 + 5,769 = 8,021$

$$p = .293, n = 8,021, \quad \frac{z}{2} = 1.96$$

$$Standard\ deviation = p(1-p) = \big(.293(1-.293)\big) = .4551$$

Table 1. MOE selected and n values obtained for defining a sample size

MOE	n=(1.96)*SDMOE	n
0.010	90.08	7957
0.011	89.19	7800
0.012	75.07	5525
0.015	60.05	3607
0.050	18.02	325

sorry, I'm not following what you're doing here.

Eg. you haven't shown the calcuation leading to the n's here (in either column)

From the table above, the minimum sample size for a margin of error = 0.05 is 325 students enrolled in CIT.

???

Sampling Error

This survey could encounter coverage error because the registrar's records are incomplete. The target population coverage depends on the completeness of the registrar office records. When a student leaves on-campus housing, they are asked to update their address on SIO but we suspect that many fail to do so. Additionally, some people may change addresses multiple times and fail to update their information. One solution to this problem would be to find the ratio of current students living in on-campus housing and weight our sample to account for any discrepancies. We could easily find the correct ratio by dividing the number of students living in dorms by the total student body.

I don't see how this solves the multiple addresses problem.

This is a coverage error issue, not a sampling error issue.

Data Collection

We collected data from administrative records provided by the office of the registrar. We believe surveying data records is a more accurate and reliable method in comparison to asking students directly. This mode of collection and survey can help reduce high non-response and coverage errors.

We were successfully able to ~~attain~~ obtain off-campus housing records from the University registrar. The records have 891 undergraduate records and 4,036 graduate records. The registrar provided us with all the records that they had. According to the Factbook, there are 2,252 undergraduates living off-campus and 5,769 graduates living off-campus. Clearly, the ratio of undergraduate records to graduate records is not the same as the population ratio, but this could be explained by response errors more relevant to undergraduates.

As I have said before I want you to treat the data you got from the registrar as a (post) stratified sample from the full population of off-campus students.

Analyses within a stratum can be done using SRS techniques. Analyses that combine the two strata should use post-stratification weights, etc.

Possible reasons and sources for apparent bias

Most undergraduates start their CMU careers living on-campus so changing their address to an off-campus location will probably be less likely reported to the registrar (especially if they still use their SMC mailboxes to get mail from the university). Other sources of bias in the collection of data could be the limitation of department information. When looking at clusters of students off-campus according to their major, a student could have more than one major, but the records only indicate one major and one affiliated department per student. Another possible bias is that students may not have reported accurate addresses of zip codes e.g. using abbreviations or interchangeable zip codes. We had to sort through the data to locate these inconsistencies as part of the data cleaning process.

Given that we have obtained all of the records from the registrar for students living off-campus that provided responses, we are going to analyze our sample with two different methodologies: as a census and as a stratified sample. As part of cleaning the data, we noticed that graduate students have a duplicate entry for their offices, therefore, we made sure to only report their residences in our results. Other issues we needed to consider when cleaning the data were duplicate records, response missingness, and incorrect forms of address format.

Questionnaire

In general, the questions included in the survey consisted of which department, school, and class year the student belonged to as well as where the student lived and the distance and time it took to travel to campus from their off-campus residence.

A sample of questions included:

- Identification of class
    - Does this record belong to an undergraduate student?
    - Does this record belong to a graduate (Master) student?
- Identification of college/department
    - Does this record belong to a student enrolled in the School of Computer Science (SCS)?
        - Which department?
    - Does this record belong to a student enrolled in the College of Fine Arts (CFA)?
        - Which department?

please include complete set of questions and/or variables and their definitions, in an appendix to this report

Post-Survey

Based on all the data we obtained from the office of the registrar, we had to format our data into a uniform coding system so that it could be used for the analysis. After reviewing the data, we found that we had to omit 182 records due to 157 students reporting campus addresses, 16 reporting duplicate addresses, six students with incomplete addresses, and three reporting no addresses to the office of the registrar. We did not include these records in our dataset since our question of interest is related to only assessing off-campus housing for students. We decided that for students who reported two addresses, we would use the first address listed to be included in our dataset. This way we were still able to include one of the addresses provided and just omit the other address from the dataset. We found that we had to re-format addresses to move forward in the analysis.

how many of these were grad, how many undergrad? what did that leave for final sample sizes?

Our main variables included in the current analysis are address, class year, college, department, and distance to campus. The housing variable was comprised of our address list and was coded into street name, apartment, city, and zip code. To estimate the distance to campus for our distance variable, we used the ArcMap Geographic Information Systems (GIS). Using GIS we were able to obtain a map of all the addresses within the city of Pittsburgh and use it to estimate the distances to campus. The class year variable was separated into undergraduate, master's, and PhD students categories. The college variable was comprised of eight distinct colleges which included CFA, CIT, CMU, HC, HSS, MCS, SCS, and TSB. The department variable included 64 distinct departments listed.

Many addresses listed included students who live in cities that are outside the city of Pittsburgh, such as Homestead, Monroeville, etc. The following analyses are based on a total of 3,888 addresses that were only in the city of Pittsburgh. However we plan to explore if we can use the 202 other addresses found in neighboring cities for our analysis once we find a way to add separate maps of these cities to the Pittsburgh map and can estimate their distances using GIS.

Results

Class year, college, and department variables were coded into R to obtain demographic information about the students. Of the eight colleges represented, CFA had 398 students, CIT had 1,275 students, CMU had 147 students, HC had 453 students, HSS had 318 students, MCS had 271 students, SCS had 632 students, and TSB had 394 students. We found that 20.3% of the students living off-campus were undergraduate students, 47.1% were Master's students, and 32.5% were PhD students. The distribution of college and class year is shown below in Figure 1 and Figure 2, respectively.

these are the sorts of things you can form CIs for, for the full target population of off-campus students, by treating your registrar data as a stratified sample.

note that you cannot do this for percent undergrad, say, since undergrad vs grad is your stratifying variable.

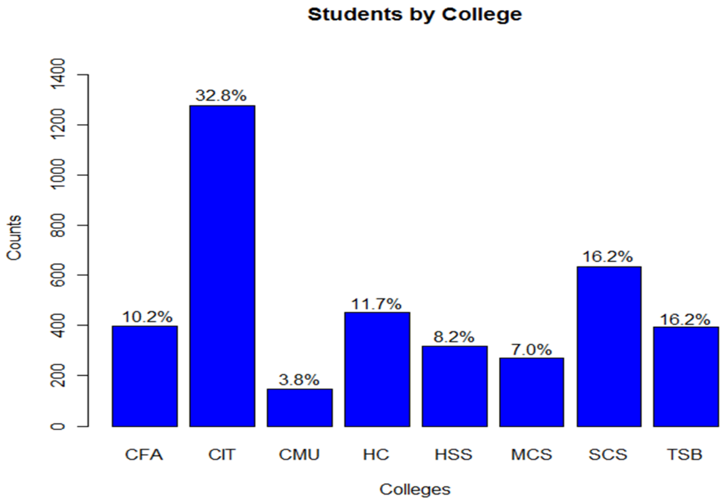


Figure 1: Showing students living off-campus by college.

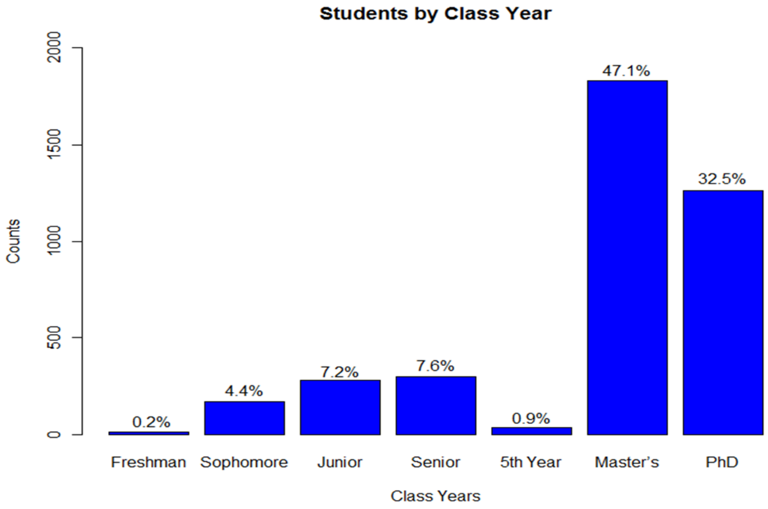


Figure 2: Showing students living off-campus by class year.

As for the departments represented by the students, we found that the highest numbers of students living-off campus were from the departments of Electrical and Computer Engineering, Industrial Administration, and Mechanical Engineering with 410, 333, and 227 students respectively.

ok

We are still working on the analysis (both visual, e.g. maps, and statistical). However, to show our progress, we are planning to include variations of the tables and maps seen below. Also we are working on the table with the post-stratification weights. We plan on doing statistical analysis of the distances (between residences and campus, shuttle stops, etc.) and compare this analysis between the census sample and the stratified sample. We encountered some issues getting the spreadsheet of distances from GIS, but evidently, it is a problem that a little bit of time (and fiddling) can fix. Most of our analysis is based on these numbers.

Table 1: Representativeness of our data vs. CMU's full records (by college)

Distribution by College

College	Undergraduate		Master's		Doctoral	
	CMU	Our data	CMU	Our data	CMU	Our data
CFA	77.3%	51.1%	21.0%	43.9%	1.7%	5.0%
CIT	52.9%	16.8%	24.5%	43.4%	22.6%	39.8%
DC (HSS)	79.1%	49.4%	6.7%	13.1%	14.1%	37.5%
HC	0.0%	0.0%	96.1%	92.8%	3.9%	7.2%
Interdisc. (CMU)	58.8%	26.5%	41.2%	73.5%	0.0%	0.0%
MCS	71.2%	29.7%	2.3%	5.4%	26.5%	64.9%
SCS	40.7%	11.2%	28.6%	37.7%	30.7%	51.1%
TSB	28.2%	13.0%	63.9%	66.0%	7.9%	21.0%

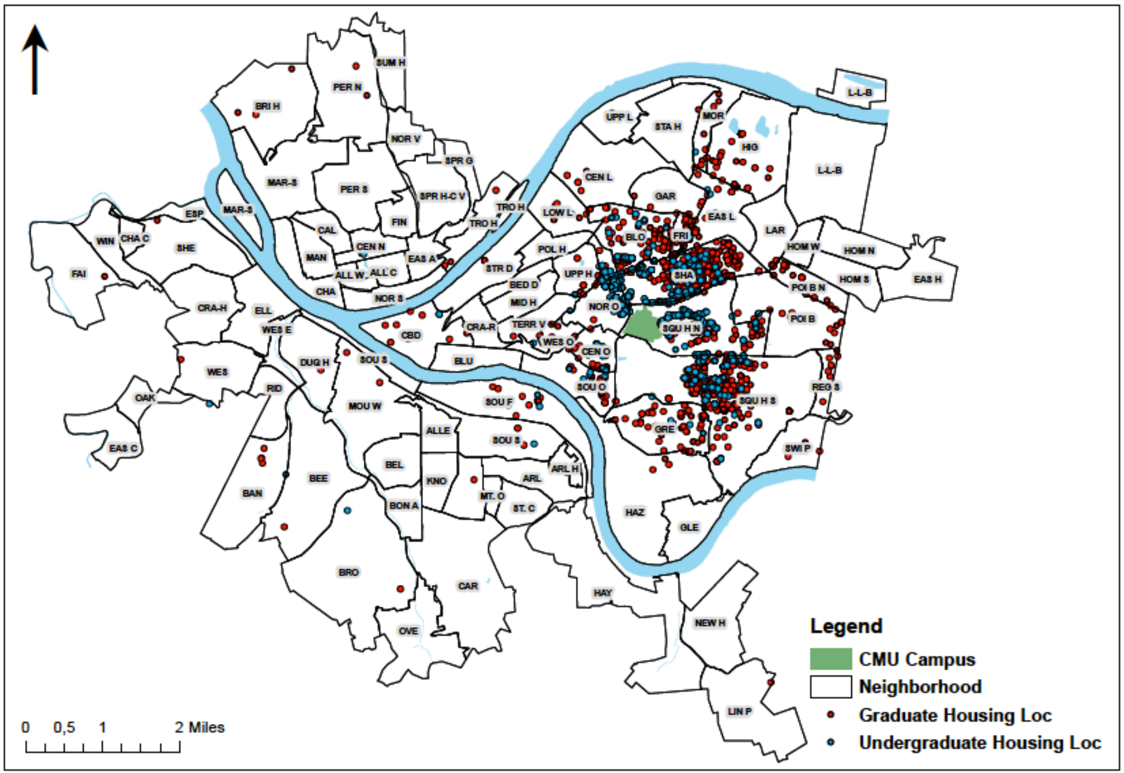
Table 2: Representativeness of our data vs. CMU's full records (by class)

Distribution by Class

College	Undergraduate		Master's		Doctoral	
	CMU	Our data	CMU	Our data	CMU	Our data
CFA	16.8%	24.6%	6.9%	9.6%	1.2%	1.6%
CIT	30.1%	25.7%	21.3%	30.2%	40.7%	39.7%
DC (HSS)	20.1%	20.0%	2.6%	2.4%	11.4%	9.9%
HC	0.0%	0.0%	31.9%	23.6%	2.7%	2.6%
Interdisc. (CMU)	4.6%	4.7%	4.9%	5.9%	0.0%	0.0%
MCS	12.4%	10.1%	0.6%	0.8%	14.7%	14.4%
SCS	10.0%	8.5%	10.7%	13.0%	23.9%	25.2%
TSB	6.1%	6.3%	21.2%	14.6%	5.5%	6.7%

since the CMU data is for the entire CMU student population, these tables show some interesting differences between those who chose to live off campus and those who stay on campus. What other demographic differences can you find between these two populations?

Figure 3: Graduate vs. Undergraduate Students Housing Location



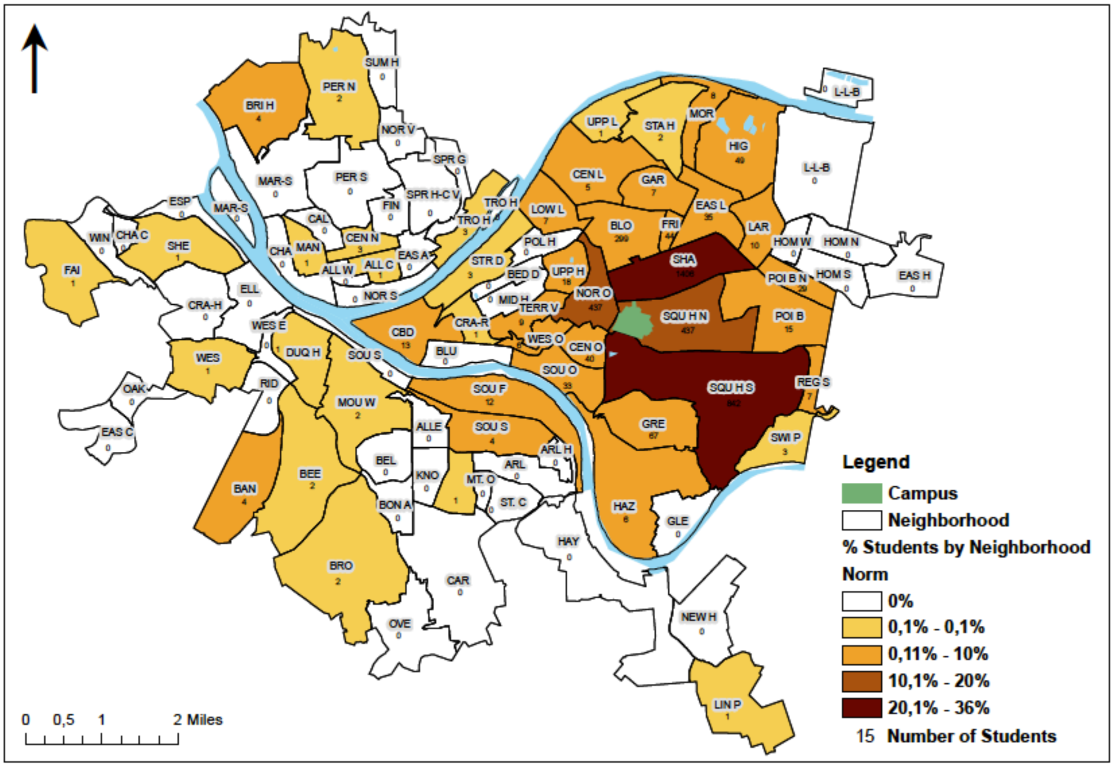
Source: CMU Registrar's Office

these will be interesting.

in much the same way that we talked about adjusting boxplots and histograms for weights in class, you may be able to adjust these plots for weights (since they are basically two-dimensional histograms).

that would begin to show how the full target population is distributed across the pgh area.

Figure 4: Distribution of All Students - Percentage and Number by Neighborhood



Source: CMU Registrar's Office

Discussion

We will be able to finish this section when our analysis is complete!!

List of References

Fitzgerald, M. R. (2006, September 25). Housing: A Financial Look. *The Tartan*.

Office of Institutional Research and Analysis. (2012). Carnegie Mellon University Factbook 2011-2012. *26*. Pittsburgh: Carnegie Mellon University. Retrieved April 2012, from http://www.cmu.edu/ira/factbook/facts2012.html

Pascarella, E., & Terenzini, P. (2005). *How College Affects Students: A Third Decade of Research (Vol. 2).* San Francisco: Jossey-Bass: A Wiley Imprint.

Schudde, L. T. (2011, Summer). The Causal Effect of Campus Residency on College Student Retention. *The Review of Higher Education, 34*(4), 581-610.

Snyder, T., Dillow, S., & Hoffman, C. (2009). *Digest of Education Statistics, 2008.* Washington, D.C.: National Center for Education Statistics.

Yeung, R. (2011). *A Quasi-Experimental Approach to Estimating the Impact of College Housing*. Syracuse, NY: Syracuse University.

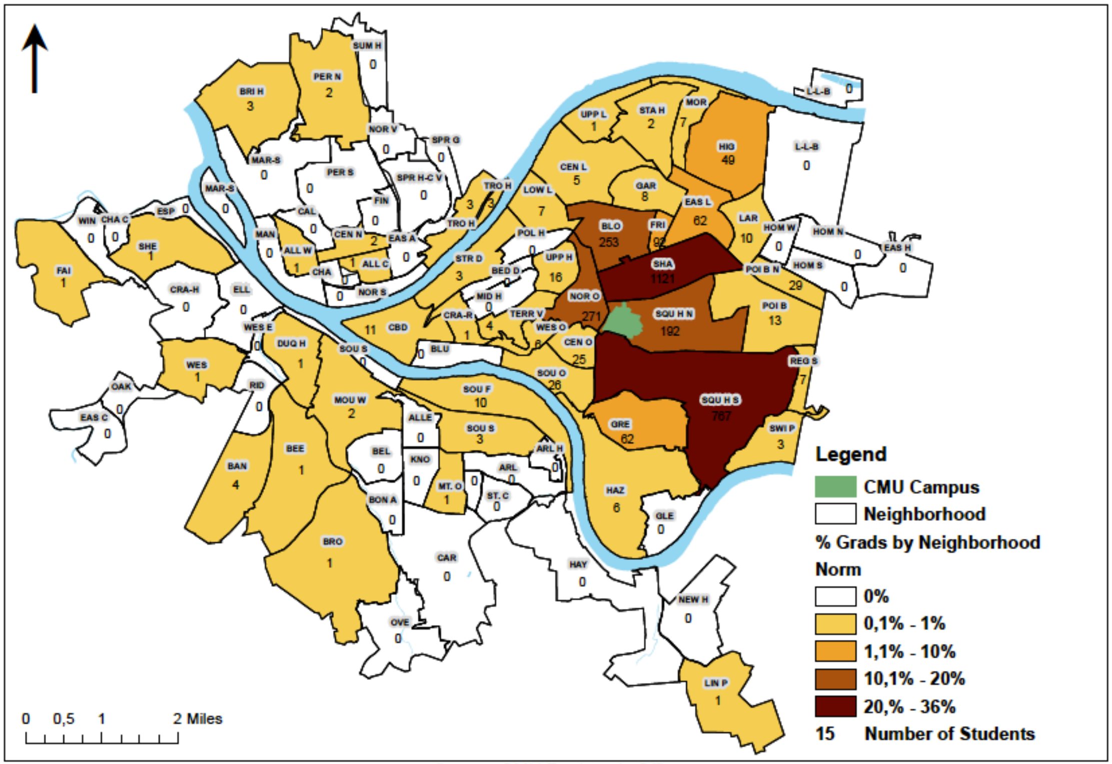
Appendix A

We will include most of our tables (and gory statistics work) in the appendices. Any of the data analysis that we feel is important to our paper, but is not necessarily relevant to include in the paper will be put here. For example, we included one of the tables we had used in our progress report presentation. We also will include maps in this section.

Table A-1. Student Records Distribution vs Student CMUs Distribution

	Cleanned Records				CMU Distribution 2011/12	
	Undergr	Grad	Total	% of total	Number	% of total
CIT	224	1111	1335	32	3,217	30
SCS	74	586	660	16	1,366	13
HC	0	488	488	12	951	9
TSB	55	369	424	10	1,126	10
CFA	214	205	419	10	1,267	12
HSS (DC)	174	178	352	9	1,481	14
MCS	88	208	296	7	1,019	9
CMU	41	114	155	4	423	4
Total	**870**	**3259**	**4129**	**100**	**10,850**	**100**

Figure A-1. Graduate Students Distribution - Percentage and Number by Neighborhood



Source: CMU Registrar's Office