Mapping the Physical Internet: Where Your Data Lives

Peter Swigert, UC Berkeley School of Information
December, 2014

They want to deliver vast amounts of information over the internet. And again, the internet is not something you just dump something on. It's not a truck.

It's a series of tubes.

Senator Ted Stevens, Chair of the Commerce, Science & Transportation Committee, 2006[i]

Introduction

When an email is sent from a laptop, the data from that message goes through your laptops internal wiring, gets converted in the modem and transmitted over wifi to a router, reprocessed and sent into a series of wires that connect the router to an ISP's network to the thousands of networks that make up the internet, before coming to rest in the series of wires that make up a data center. These actions may feel intangible, but they are in fact physical processes that occur in real space. Ted Stevens was mocked in 2006 for his reference to "tubes," but history has treated his goof more kindly and recognized the physical nature of the internet.

As data becomes a commodity, the information being sent through over the internet is increasingly being stored in permanent data repositories. As the internet connectivity becomes increasingly central to our lives and digitally captured data drives increasing numbers of decisions, it is helpful to understand where the internet actually is, where the data that is shared over the internet actually lives, and why it lives there.

The movement and storage of data have significant implications for policymakers in a number of fields, including privacy, security, energy and free speech.

- In the United States, the continuing fallout from Edward Snowden's revelations has demonstrated how the National Security Agency and other government agencies have made physical connections at key points in internet infrastructure.
- For leaders and regulators in energy, demand for internet services has 'produced a collective electricity demand that would currently rank in the top six if compared alongside countries; that electricity demand is expected to increase by 60% or more by 2020 as the online population and our reliance on the internet steadily increase.ii

- Recent legal decisions in the European Union, including the EU's 'right to be forgotten' [iii] ruling have suggested that information stored about specific individuals, should be removable at request of the user at hand. It is critical to understand where that data might be physically located, and under what jurisdictions and frameworks it might be treated if it left its permanent repository and was shared elsewhere (for instance, to a personal laptop in Europe upon a browser request).

This report does not specifically address these policy questions, but rather provides an exploratory data analysis that establishes a basic framework for understanding where data centers are located today and, at the aggregate level, what factors have influenced this distribution.

Most large sets of data are stored in facilities specifically designed as data centers, which house a few thousand servers to more than a million servers. These data centers may only serve the needs of one company who owns and operates it (as in the case of a Google data center) or may lease space and/or hardware to anyone who needs it (such as colocation providers like Equinix). A typical data center may easily cost tens of millions to build and equip with rows of racks, each housing servers, switches and routers. Siting a data center development "involves many important considerations, including its proximity to population centers, power plants, and network backbones, the source of the electricity in the region, the electricity, land, and water prices at the location, and the average temperatures at the location."[iv] Additional factors range from local tax breaks to specific workforce requirements.

However, three meta-factors that influence all storage decisions about data are:

- Access to other networks (to quickly send and receive messages)
- Access to customers (to limit the physical distance and number of networks messages must traverse to reach a user)
- Access to electricity (as "a single data center can take more power than a medium-size town")[v]

Data center owners and developers have existing models to site new data centers according to business needs. However, stakeholders outside of the business lack insight into where data centers and other internet infrastructure may be developed. By providing a high level analysis of these three factors it is possible to identify patterns in data center distribution that may help future policymakers understand what the implications of physical internet infrastructure may be for their domain.

Methodology

Data Collection

Many owners and operators of data centers are hesitant to publically share extensive information about their facilities. Additionally, data centers are being constructed and renovated at a rapid rate. As such, there are no comprehensive data set of data center locations and attributes. Due to this limitation, a sample of 127 data center locations was scraped from corporate websites and technology publications to cover companies of public interest and a mix of market leaders in owner/operator facilities with a single tenant (Google, Microsoft, Facebook and Amazon) and colocation (Equinix, CoreSite, Sabey, Verizon and AT&T) facilities. Latitude and longitude were assigned by looking up specific addresses. When data center addresses were not available (~50% of the data set), city centers (as determined by Google Maps) were assigned. Regarding attribute data, data center owners rarely publish statistics on number of servers, storage or processing capabilities of data centers. Square footage was initially used as a weighting factor, but numbers were not available across all data centers, were dependent on a wide variety of sources and often did not specify if the value was for square footage of rack space or total facility space.

Internet exchange points were captured from industry publication datacentermap.com. [vi] Addresses were geocoded to latitude and longitude. Undersea cables were downloaded from a previous amateur attempt to harvest all publically available undersea cable locations.[vii] While these data sources are not authoritative, they provide a starting point for this analysis. A formal study of this component of the infrastructure would rely on deeper collaboration with TeleGeography, the industry leader in monitoring, analyzing and mapping internet infrastructure.

State energy production and prices were obtained from the US Energy Information Agency (EIA) and are recent as of December 2013.[viii]

Population density data was pulled from the Socioeconomic Data and Applications Center (SEDAC) at Columbia University [ix] . 2015 population projections were used, as many data centers in the data set have been built or upgraded extensively within the past five years and are generally built to account for future capacity.

SEDAC Population Density

Data Analysis

All files were processed in Python, with maps generated via Python libraries or QGIS. The primary codebase can be accessed at http://nbviewer.ipython.org/gist/pete-sw/487a381811e0d05b5627.

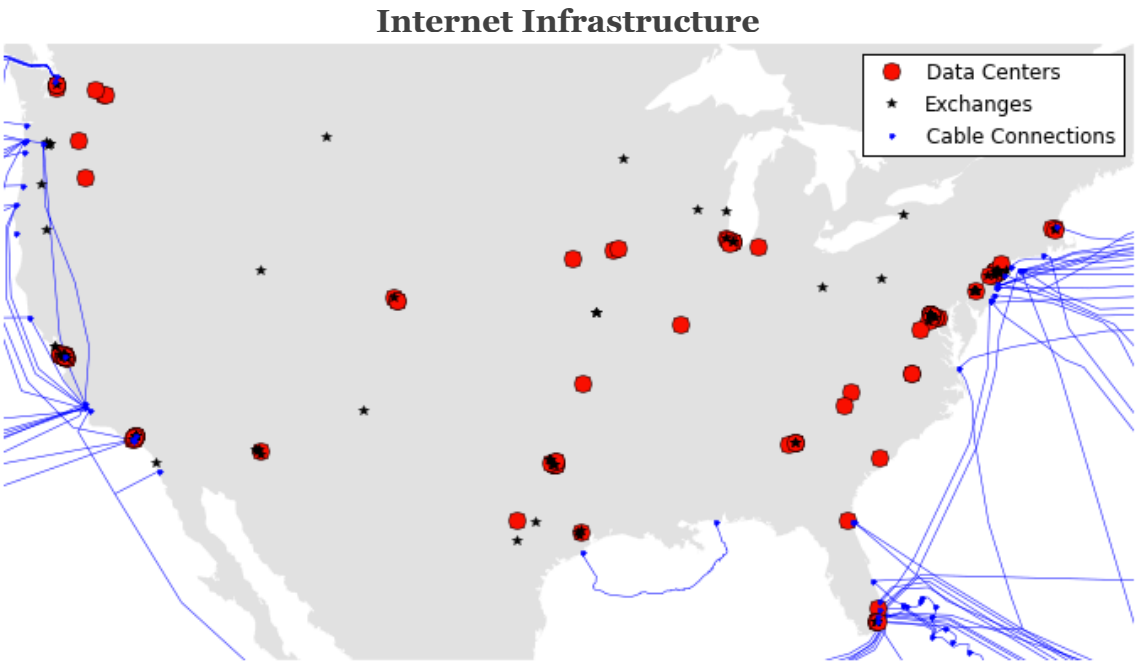
Some critical file conversions were necessary, including converting the SEDAC raster data to a vector point dataset. Points to point distances, if not otherwise specified, are calculated as if on a smooth sphere.

Data center locations showed a clear clustering pattern. Two different forms of clustering were applied. First, it was clear from initial mapping of the data that the largest clusters were in and around Silicon Valley, Northern Virginia, and New York City. Unsurprisingly, these three clusters corresponded to specific data-intensive industries (technology, government and finance, respectively), as well as large population centers. Next, an algorithmic clustered the data was applied using Density-Based Spatial Clustering[x]. DBSCAN was selected rather than a K-means approach as K-means "minimizes variance, not geodetic distance" and the distributed nature of the dataset across the continent covered a large range of latitudes.[xi] DBSCAN was able to re-identify the same three manually recognized clusters when given reasonable parameters in terms of distance between data centers to consider and number of data centers to consider. While there has been extensive research on technology and innovation clusters, there is little specific determination at this point as to how to define a "data center" cluster. As such, DBSCAN results under other parameters were noted and deserve further exploration.

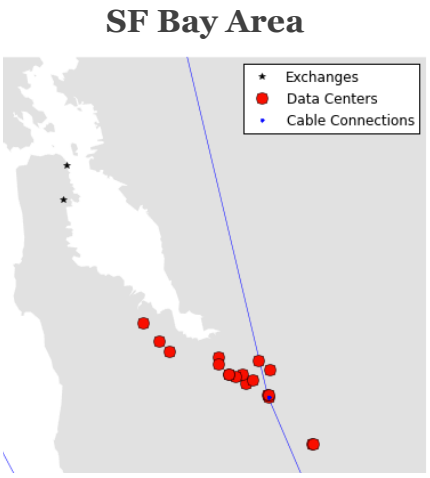
Results

Internet Infrastructure Overview

An overview map of internet infrastructure shows its distribution across the United States with clusters around major population centers, particularly along the coasts.



Most major cable links come into major ports in and around New York and Miami on the Atlantic coast, and are distributed along the Pacific coast, with a major cluster connecting in the Central California coast. Internet exchange points center are major cities, particularly New York, Washington, D.C., Chicago, Los Angeles, and New York. However, there are also individual facilities and sets of data centers distributed loosely across the West and Midwest.

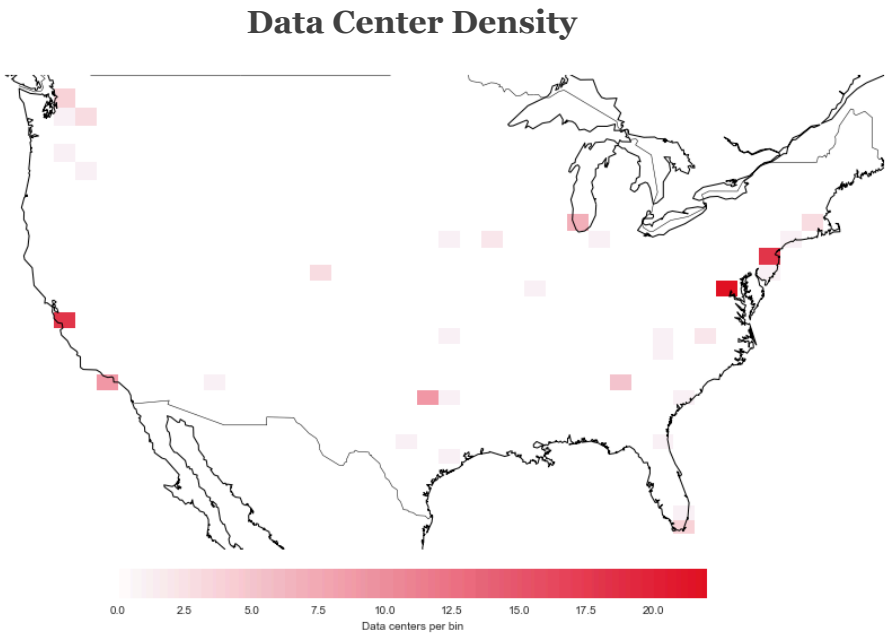


Data center locations appear tightly correlated with major internet exchange points and undersea cable connections that link networks across the Atlantic and Pacific. A close up of the San Francisco Bay Area shows the density of data centers in Silicon Valley and greater San Jose, with two internet exchange points in San Francisco.

A trans-Pacific cable has come ashore and runs North/South, connecting at an exchange point in San Jose. This suggests that that exchange points located

in San Francisco have connections to smaller cables linking fewer networks, a possible contributing factor (of the many interrelated factors) to the clustering of data centers in South Bay.

Data centers are distributed primarily along the coasts and Midwest, suggesting a strong relationship with population centers. Plotting points fails to demonstrate the density of data centers where they cluster in certain areas, but a heat map emphasizes primary clusters around Silicon Valley, New York and Washington, D.C.



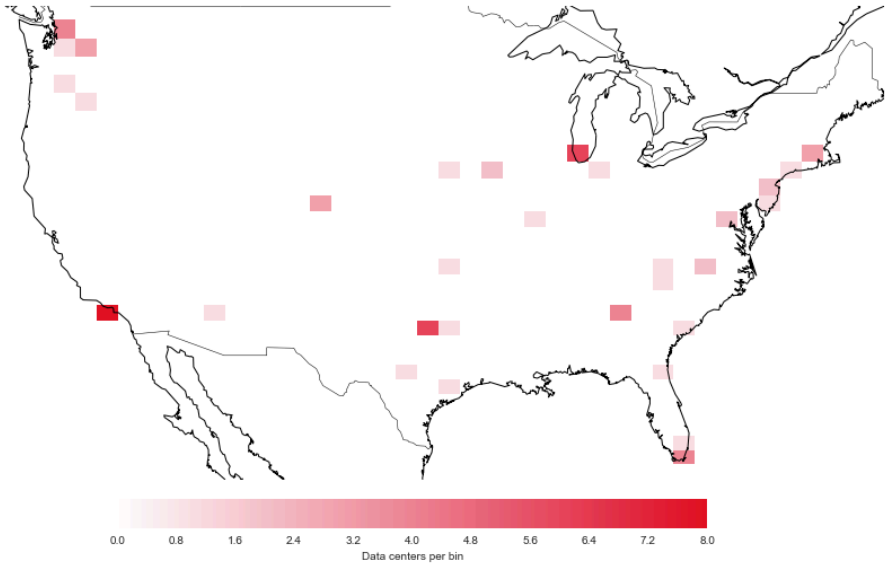
Data Center Clusters

Density-Based Spatial Clustering yielded a spectrum of results. Keeping the epsilon (maximum distance to associate two points as neighbors) constant, searching for clusters of 15, 10, 5, or 3 data centers yielded a relatively smooth increase in clusters.

Density-Based Spatial Clustering

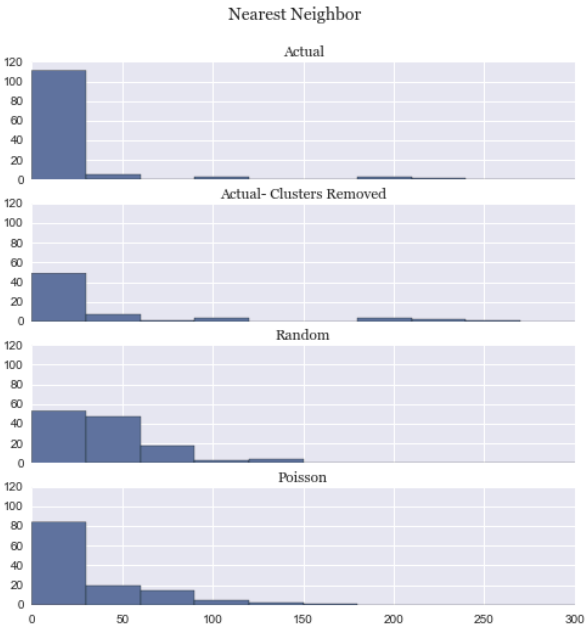
Removing the three primary clusters in Silicon Valley, New York and metropolitan Washington better captures the decreasing size of clusters.

Data Center Density (Minus Top Three Clusters)

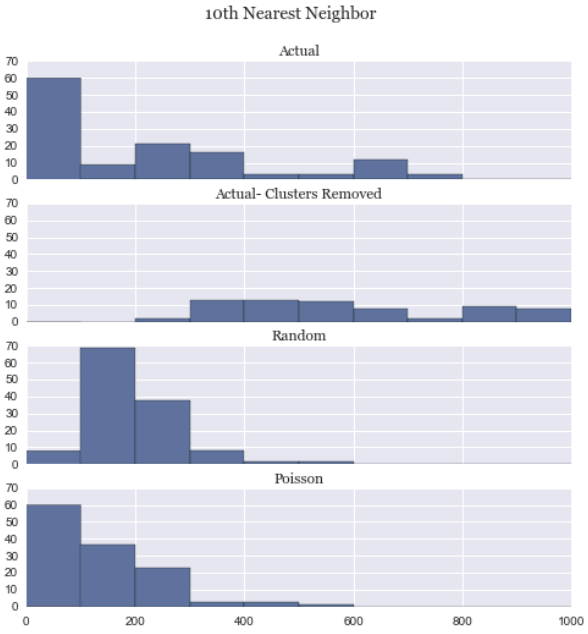


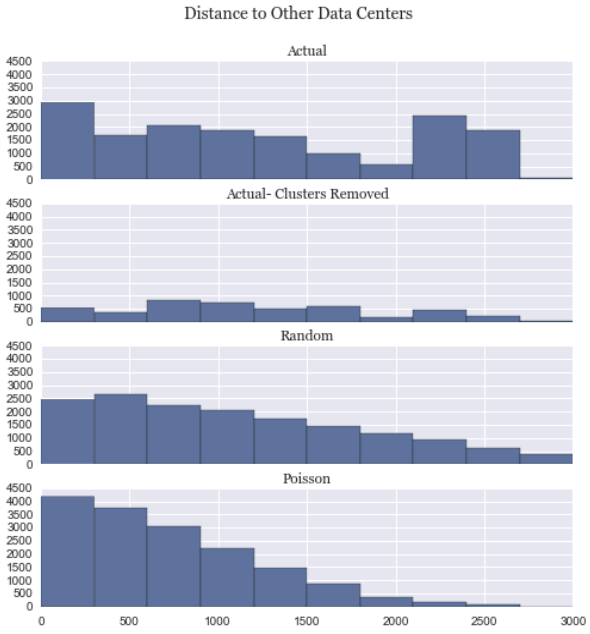
Data Center Distribution and Density

The existing set of data centers was compared to Poisson and randomly distributed sets of equal number (n = 127) data centers to compare distribution and density.



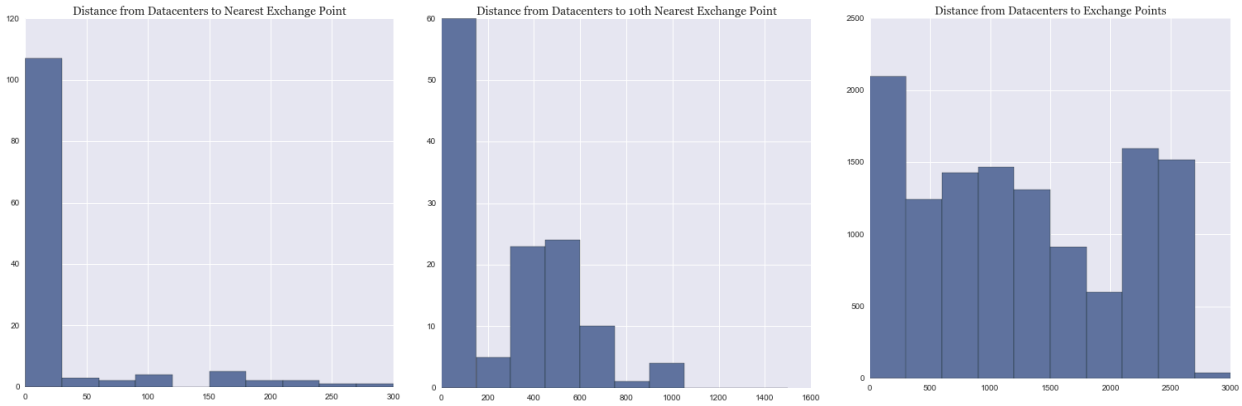
The actual distribution of data centers showed a significantly higher proportion of data centers that were located within <100 miles of the nearest neighbor, with >90% being at least that close to another data center, even with the top three clusters removed. In contrast, a random distribution across all possible longitude and latitudes, as well as a poisson distribution centered on the centroid of the U.S., showed must less clustering. 10th nearest neighbor analysis demonstrated similar results, although some rural/dispersed data centers were highlighted. In the aggregate, distance to all other data centers showed the coastal distribution of data centers as compared to more evenly dispersed random or poisson distributions.





This tracks closely with nearest neighbor, tenth nearest neighbor, and distance to all other points for exchange points.

Internet Exchange Distribution and Density

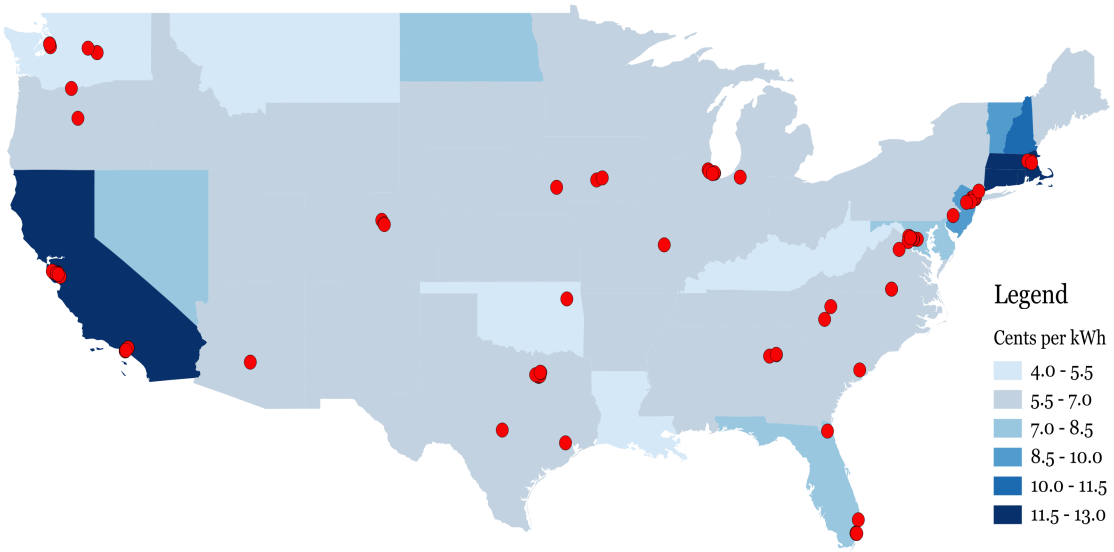


This suggests that exchange points, which are developed by a somewhat overlapping set of companies (such as telecom companies like AT&T) but not by other major players (such as Amazon or Facebook), may either be the most significant driver for data center creation or may be so similar to data centers in their requirements that both types of internet infrastructure are created in similar locations.

Energy Prices

A mapping of average industrial electricity prices by state against data center locations yields some contrasting results.

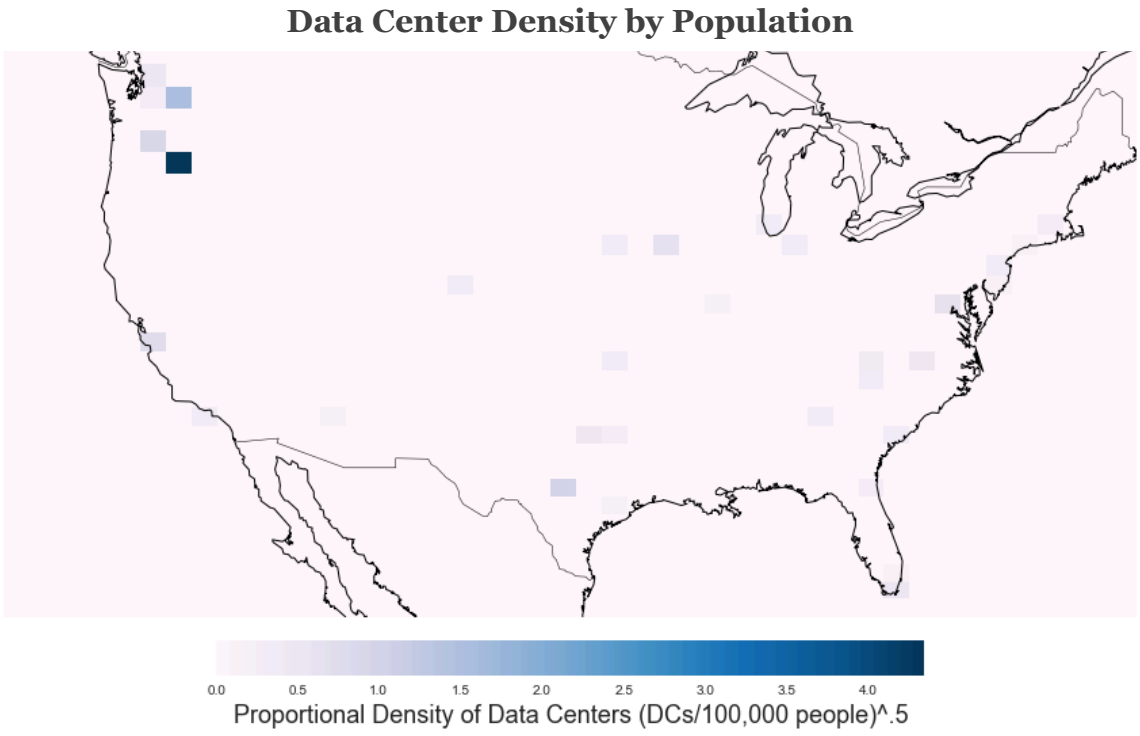
Average Industrial Electricity Prices



California, with two major clusters of data centers in Silicon Valley and Los Angeles, has some of the highest energy prices in the nation but has 26 of out the 127 data centers in the sample (.67 per million people). Massachusetts is also incredibly expensive, but has only 3 data centers (.45 per million) in the sample despite a large, densely populated city, and a strong technology industry. Boston has few cables and internet exchange points, but is as close to major undersea cable connections near New York as Silicon Valley and Los Angeles are to the onshore connections in the central coast of California. It is possible that Boston is in fact too close to New York, and that data centers would rather pay cheaper electricity costs in greater New York City and have closer access to exchange points than have a closer proximity to the metropolitan Boston population. Additionally, the difference could be driven by the clustering of technology firms in Silicon Valley and the failure of Boston's Route 128 technology corridor.[xii] Similarly, while data centers in Washington State may be attracted by cheap electricity and high level of hydropower, they may also have been developed due to existing infrastructure and worker skillsets from Microsoft, Amazon and other tech companies headquartered near Seattle.

Population

As shown in the sparse map below, rural Washington and rural Oregon are extremes in proportional number of data centers by population. These locations appear to be driven by the strong combination of cheap (often renewable) electricity, access to trans-Pacific cables and a strong network of exchange points, and possibly proximity to technology firms and an employee base in Seattle. As normalized for population, the previous clusters in Silicon Valley and the New York and D.C. metropolitan areas appear closer in scale to clusters in cities like Chicago and Dallas. However, the square root adjustment is required to effectively scale and visually capture the full range.

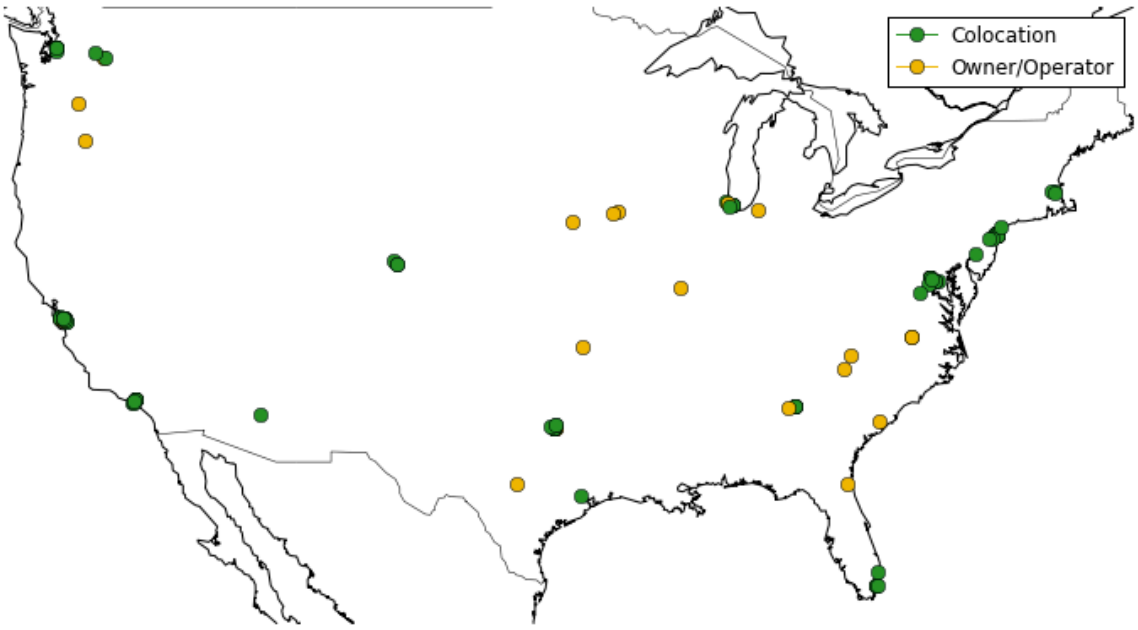


This analysis in particular highlights the challenges of the limited sample of data centers. With only a handful of major metropolitan areas represented by at least a few data centers, an adjustment of just a couple more data centers to a city like Atlanta, Dallas or Chicago could have a significant impact in the conclusions that could be drawn from this population correlation.

Colocation vs. Owner/Operator Data Centers

While major clusters have significant numbers of both colocation and owner/operator data centers, there appears to be a slight tendency towards dispersed, rural facilities to be owner/operator centers. With less than 20 data centers outside of a cluster of data centers near a major metropolitan area, there is not a significant sample size to analyze this finding. However, it is possible that different business requirements for colocation facilities limit their ability to locate far away from population centers. For instance, they may be reliant on a subset of customers who need limited physical distance to other data centers (a common theme in New York, where millisecond delays due to longer cables prevent automated algorithms from making stock trades before other request reach a server).

Colocation vs. Owner/Operator



Conclusions

Global data center IP traffic is currently estimated at 3.8 zettabytes per year.[xiii] This gargantuan set of data exists in real, physical space.

While Ted Stevens was (somewhat) right about the internet being a series of tubes, it is also a set of boxes. Hundreds of thousands of blinking, black boxes sitting on row after row of server racks, the physical embodiment of every keystroke entered and every touchscreen swipe. Understanding where these boxes are, what they connect to, and why they were installed there is an important concept for policymakers across fields as diverse as human rights and national security.

At the very least, this exploratory data analysis demonstrates some clear patterns and opportunities for further research between data center siting and variables of interest:

- Data centers are located near population centers, and particularly those with immediate access to undersea cables and internet exchange points.
- Industry, not just population, is a powerful lever, particularly tech, finance and government. For instance, the DC metropolitan area has a much smaller population than greater Los Angeles (and Southern California is comparable to the mid-Atlantic) but DC has more data centers.
- Energy prices appear to play a role in the siting of data centers, particularly in the Pacific Northwest. However, the role of prices in the largest data center clusters near major cities appears to be minor.
- Additionally, owner operator facilities are more likely to be in rural areas than colocation facilities, possibly as they can install networks and infrastructure to support specific business needs (e.g. emphasis on storage versus access and retrieval) rather than being dependent on multiple customer needs.

Further analysis is clearly needed to identify the strength of the relationships between data center locations and population density, access to other internet infrastructure, and the electric grid. A larger dataset could facilitate a spatial regression that provides increased rigor to the variables analyzed in this analyses and could incorporate additional factors.

References

[i] Single, Ryan: Poulsen, Kevin (June 29, 2006). ”Your Own Personal Internet." Wired. Retreived 2014-12-7.

[ii] Greenpeace. Clicking Clean: How Companies are Creating the Green Internet. April 2014. <http://www.greenpeace.org/usa/Global/usa/planet3/PDFs/clickingclean.pdf> Retrieved 2014-12-7.

[iii] Scott, Mark. ‘Right to Be Forgotten’ Should Apply Worldwide, E.U. Panel Says The New York Times. < http://www.nytimes.com/2014/11/27/technology/right-to-be-forgotten-should-be-extended-beyond-europe-eu-panel-says.html> Retrieved 2014-12-7.

[iv] Goiri, I.; Kien Le; Guitart, J.; Torres, J.; Bianchini, R., "Intelligent Placement of Datacenters for Internet Services," *Distributed Computing Systems (ICDCS), 2011 31st International Conference on* , vol., no., pp.131,142, 20-24 June 2011

[v] Glanz, James. Power, Pollution and the Internet. The New York Times. September 22, 2012. < http://www.nytimes.com/2012/09/23/technology/data-centers-waste-vast-amounts-of-energy-belying-industry-image.html?pagewanted=all> Retrieved 2014-12-7.

[vi] Data Center Map. <datacentermap.com> Retrieved 2014-12-7.

[vii] Greg’s Cable Map. < http://www.cablemap.info/> Retrieved 2014-12-7.

[viii] EIA. Electric Power Monthly, February 2014. <http://www.eia.gov/electricity/monthly/current_year/february2014.pdf> and <http://www.eia.gov/electricity/monthly/epm_table_grapher.cfm?t=epmt_5_6_a> Retrieved 2014-12-7.

[ix] SEDAC. < http://sedac.ciesin.columbia.edu/data/set/gpw-v3-population-density> Retrieved 2014-12-7.

[x] DBSCAN. SciKit Learn. <http://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html> Retrieved 2014-12-8.

[xi] Boeing, Geoff. Clustering to Reduce Spatial Data Set Size. August, 2014. < http://geoffboeing.com/2014/08/clustering-to-reduce-spatial-data-set-size/> Retrieved 2014-12-7.

[xii] Saxenian, AnnoLee. Inside-Out: Regional Networks and Industrial Adaptation in Silicon Valley and Route 128. Cityscape: A Journal of Policy Development and Research, Volume 2, Number 2, May 1996.

[xiii] Cisco. Global Clound Index. <http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.html> Retrieved 2014-12-7.