Popular Vacation Locations: Clustering Instagram Images based on Geotags

Parth Sawant
pss7278

Contents

1. Introduction
2. Motivation
3. Experiment
4. Clustering Algorithms & Comparison
5. Results
6. Future Work
7. References

Motivation & Data Mining

- Geotags are point in a 2D space
- Why not use clustering as a method for recommendation

Use Cases

- Trip advisors
- Tourism Industry
- Tourism Development Authorities
- Tourists
- Marketing

Data

- Original Tuple Example: 7500 Records

Attributes	Values
Id	2267
Guid	1221295116940602625_2946202682
Link	https://www.instagram.com/p/BDy6XArKe0B/
Medialink	che_key=MTIyMTI5NTExNjk0MDYwMjYyNQ%3D%3D.2
Pubdate	4/4/2016 9:41:56 PM
Author	arnabito
Title	Hammocks are the best - Playa Punta Islita, Islita, Guanacaste, Costa Rica
description	Hammocks are the best - Playa Punta Islita, Islita, Guanacaste, Costa Rica #puravida #vacation
coords	9.852863361,-85.401925983

- Cleaned: 2621 Records

Latitude	Longitude
-0.983333	-77.8167
1.049410919	103.951192
-1.092097073	35.20549007
9.852863361	-85.401925983

Experiment

- Based on the latitude and longitude values, cluster the points to reveal some patterns in the data
- Clustering centers allow us to look at regions that are highly populated with geotags

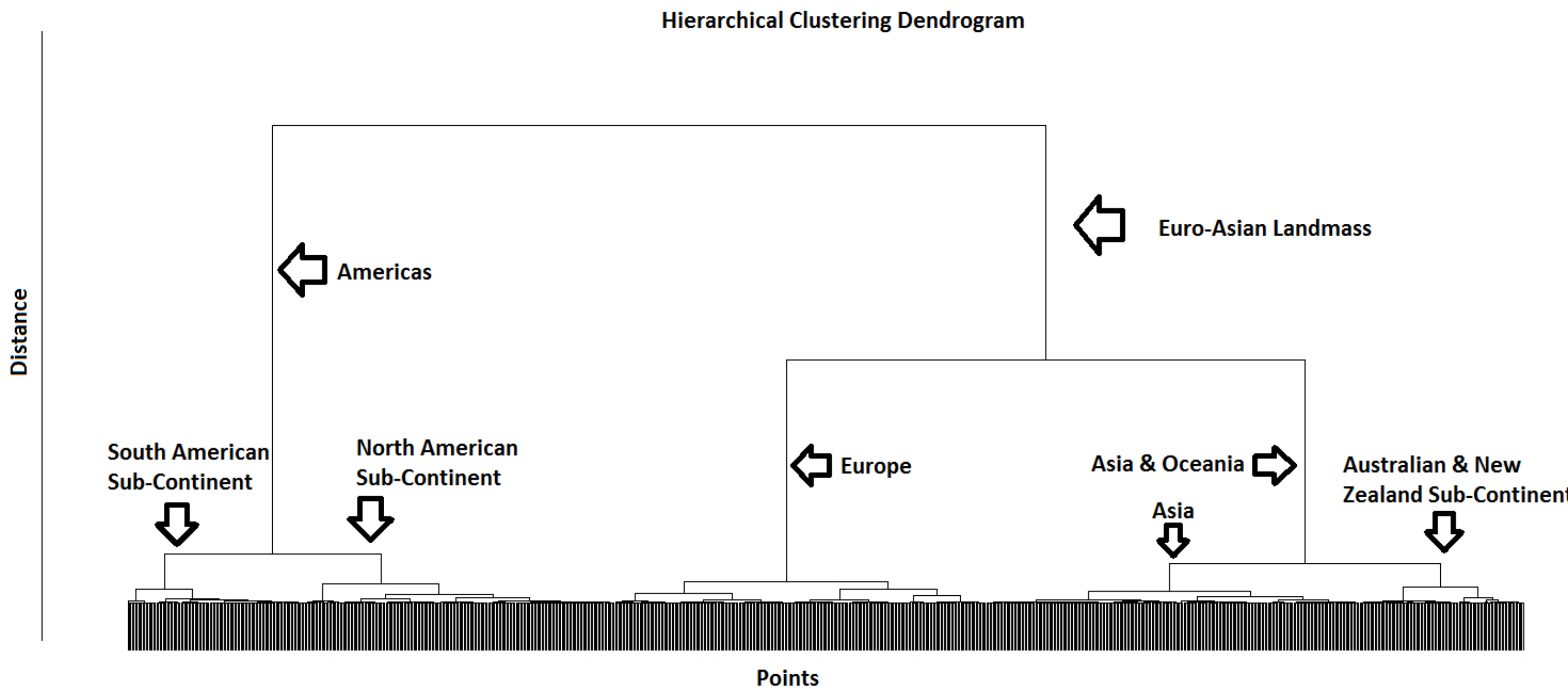
Popularity Decision

- Cluster Size a good measure of popularity?
- I chose size of cluster because its easier to explain to “Non Data Analytics People” i.e. **Stakeholders**
- Density implies small region with high number of visitors
- Population implies a highly visited area or place by many

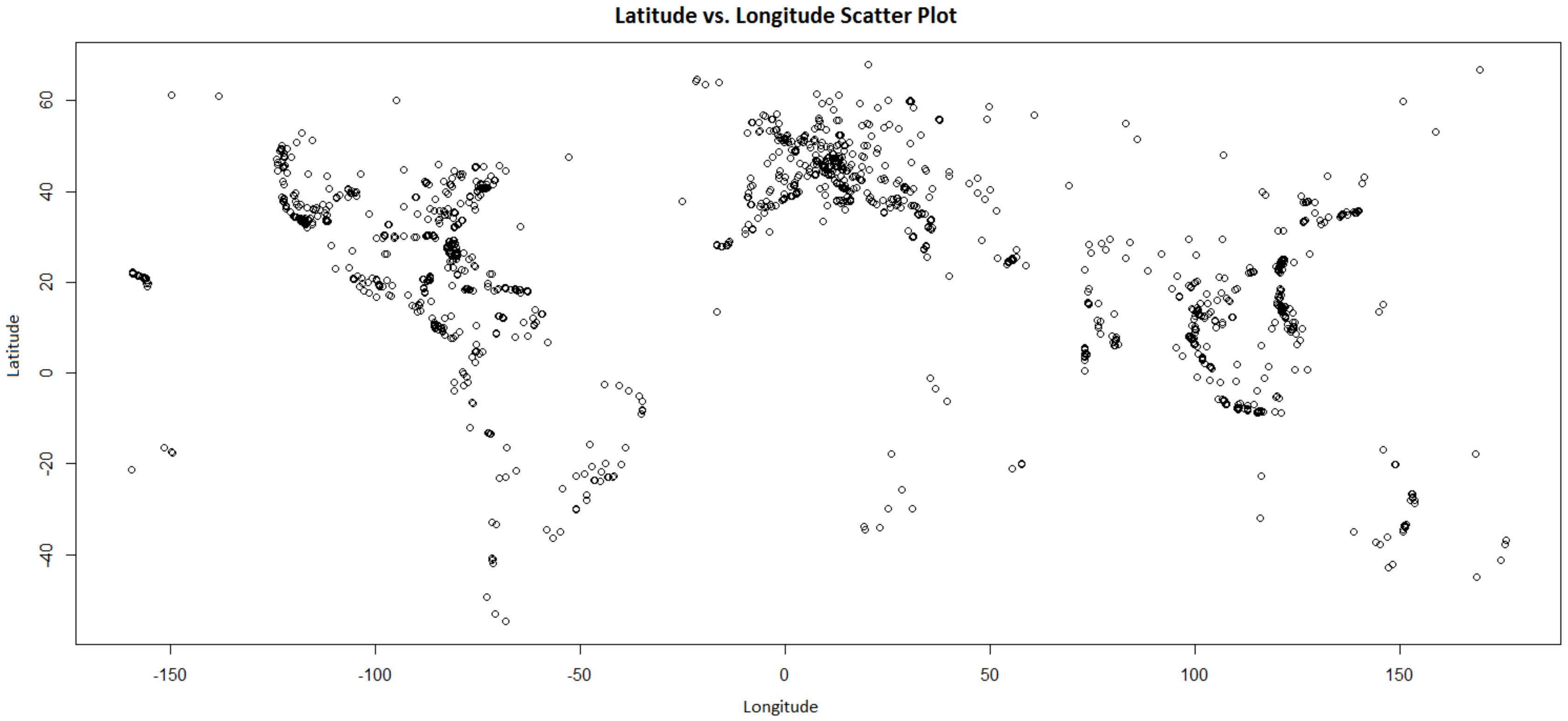
Algorithms

- K-means
- X-Means
- DBSCAN
- EM

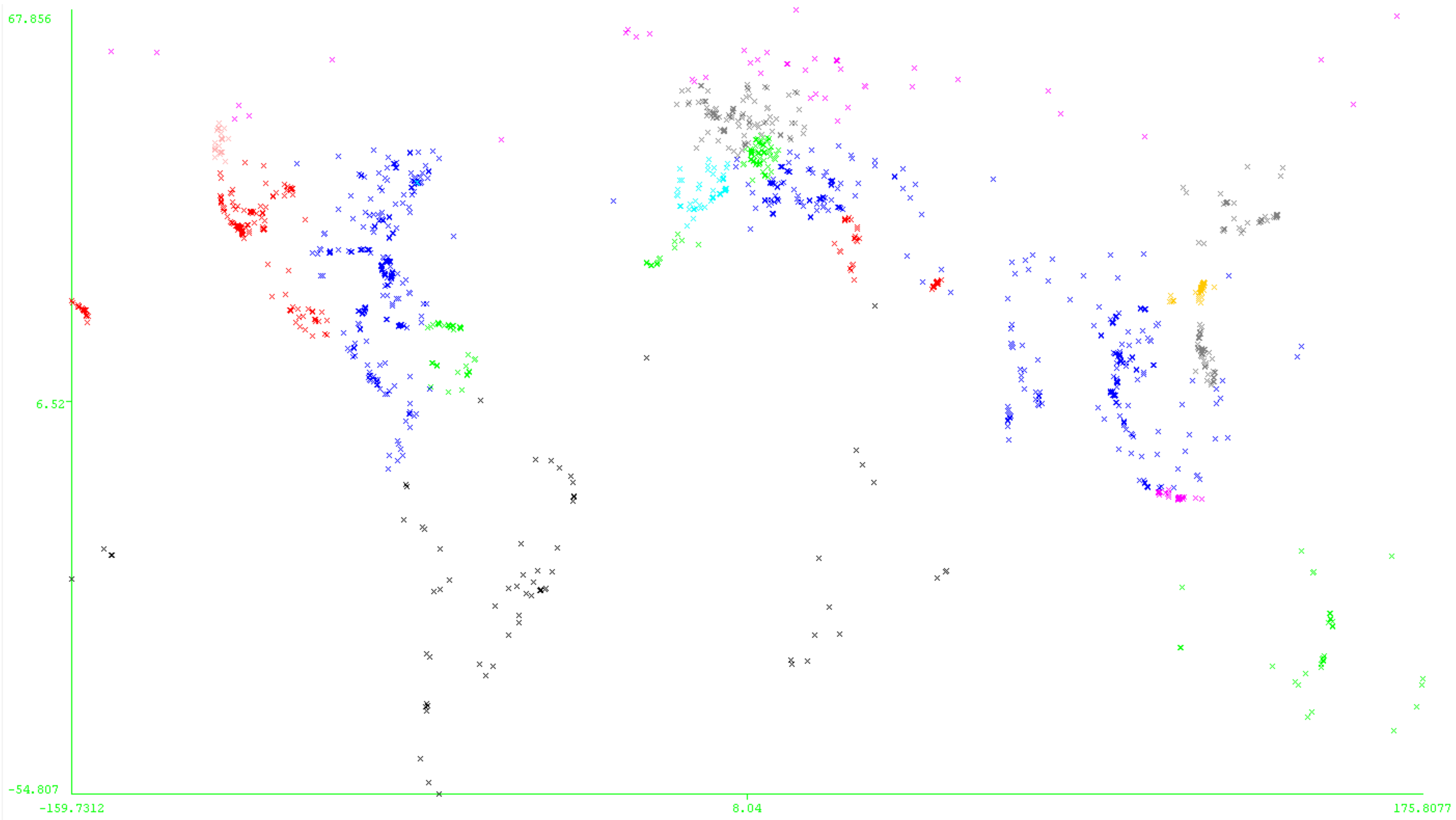
Before Results…



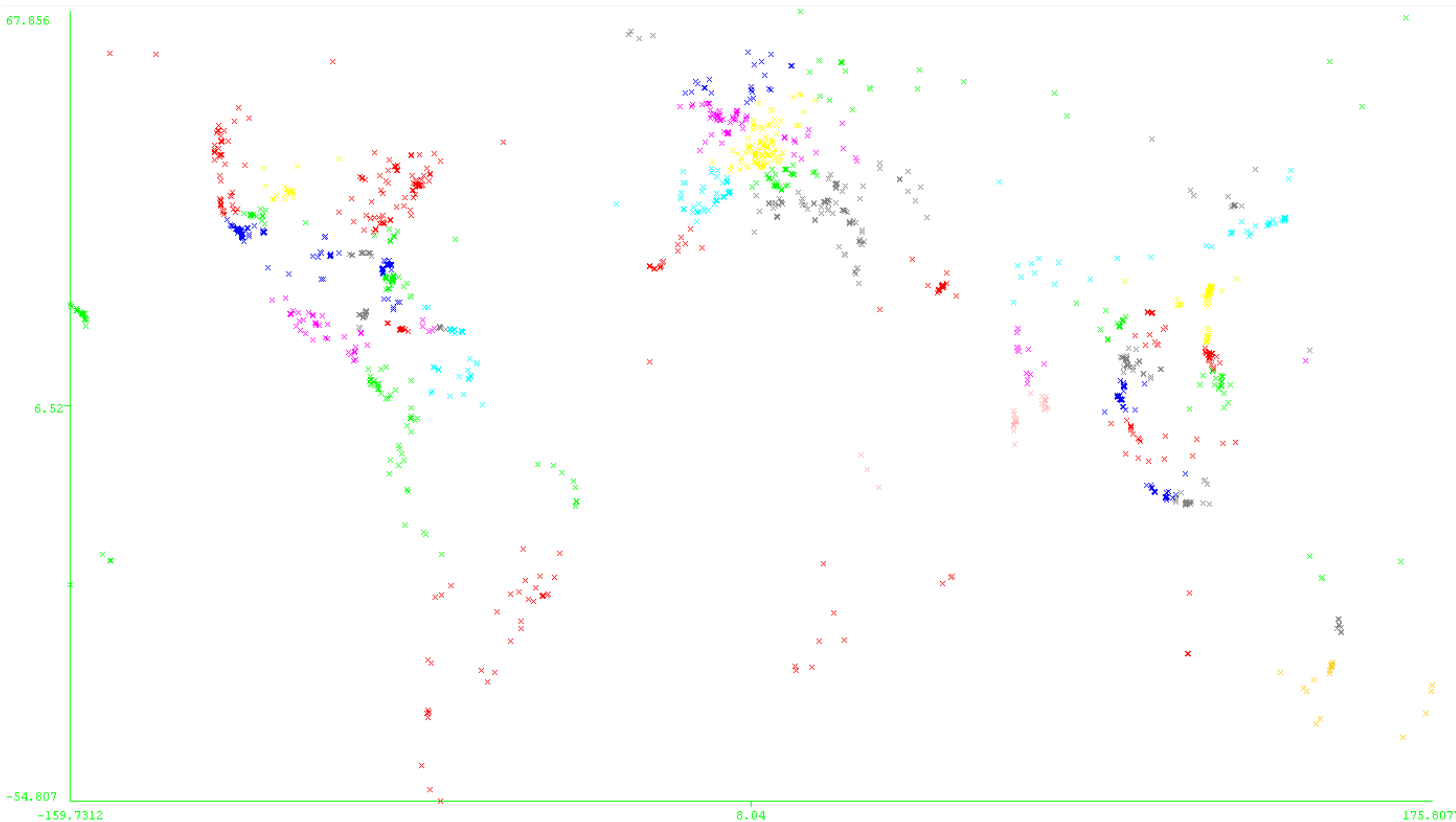
Wait another one…



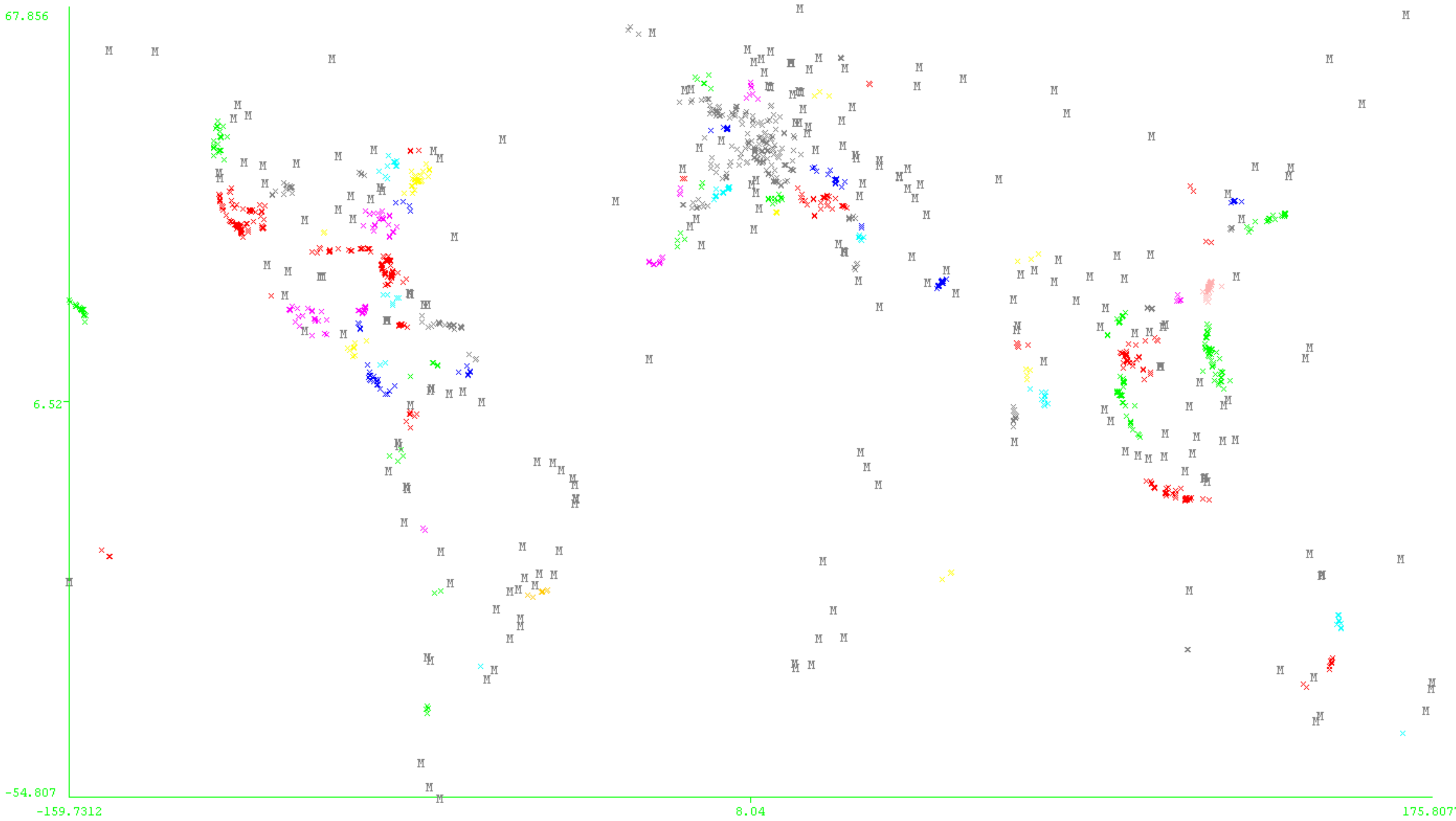
Visual: Expectation Maximization



Visual: X-Means



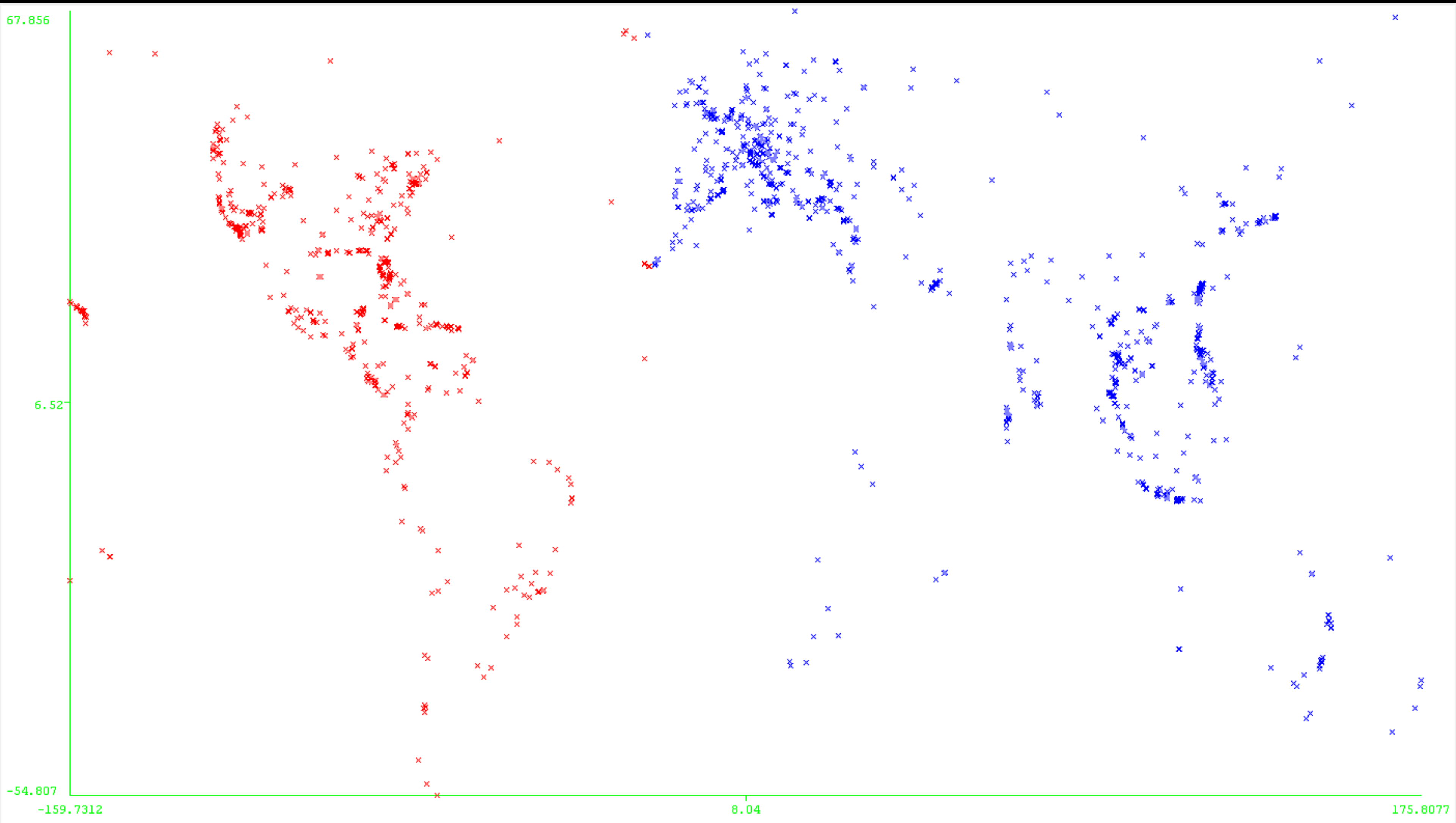
Visual: DBSCAN



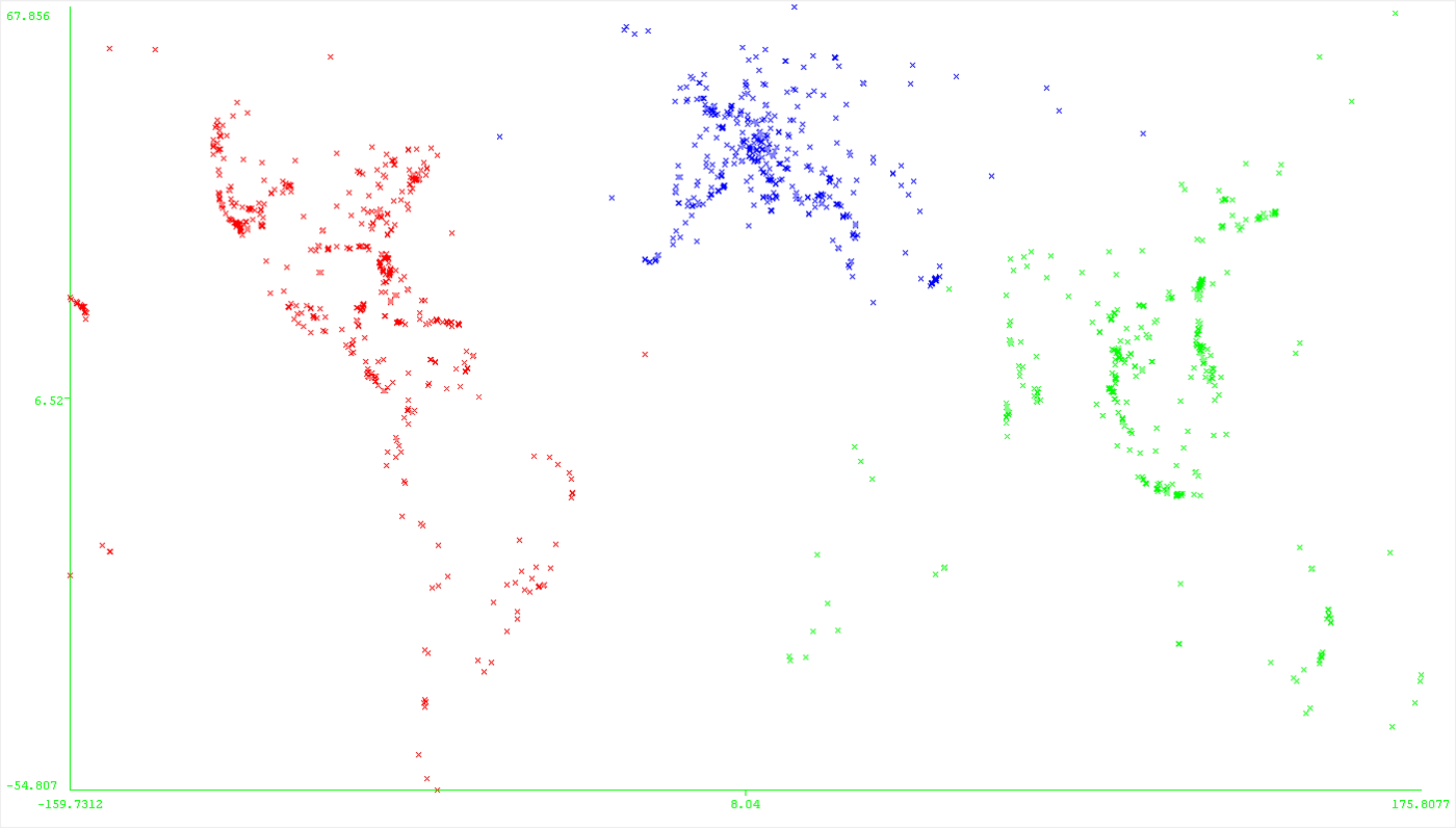
Visual: K Means

- Form k = 2
- Until   k = 65

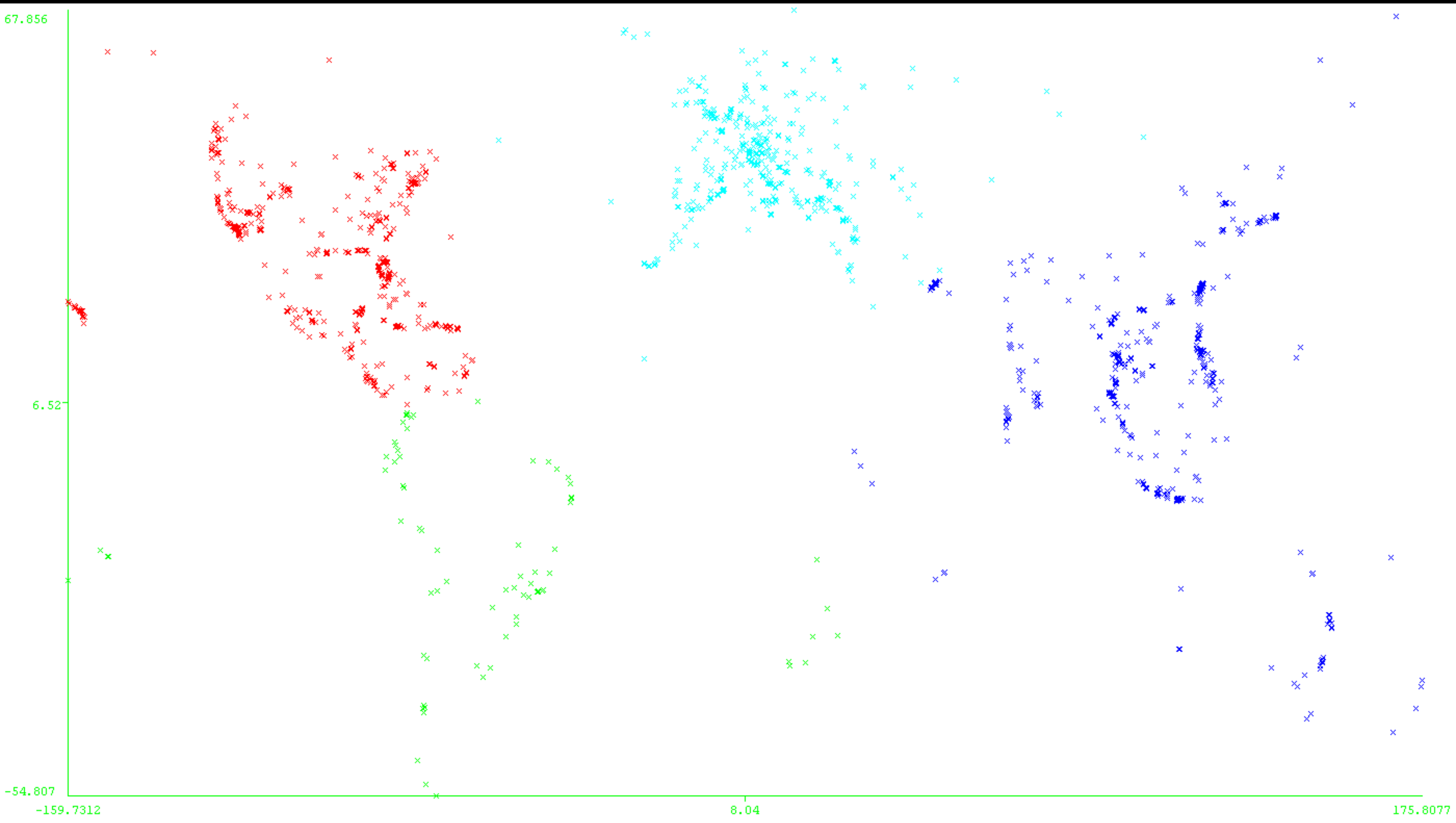
2



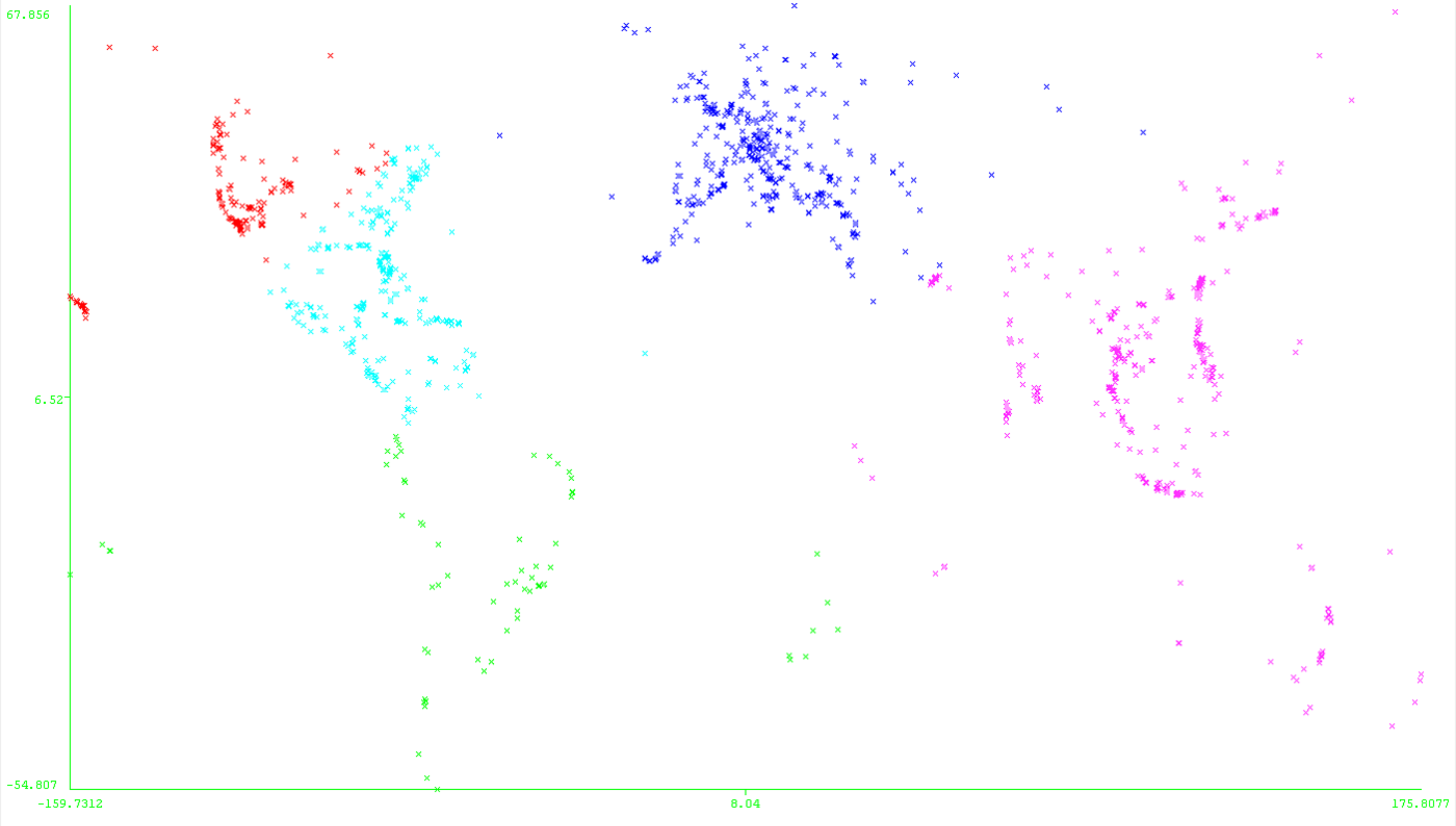
3



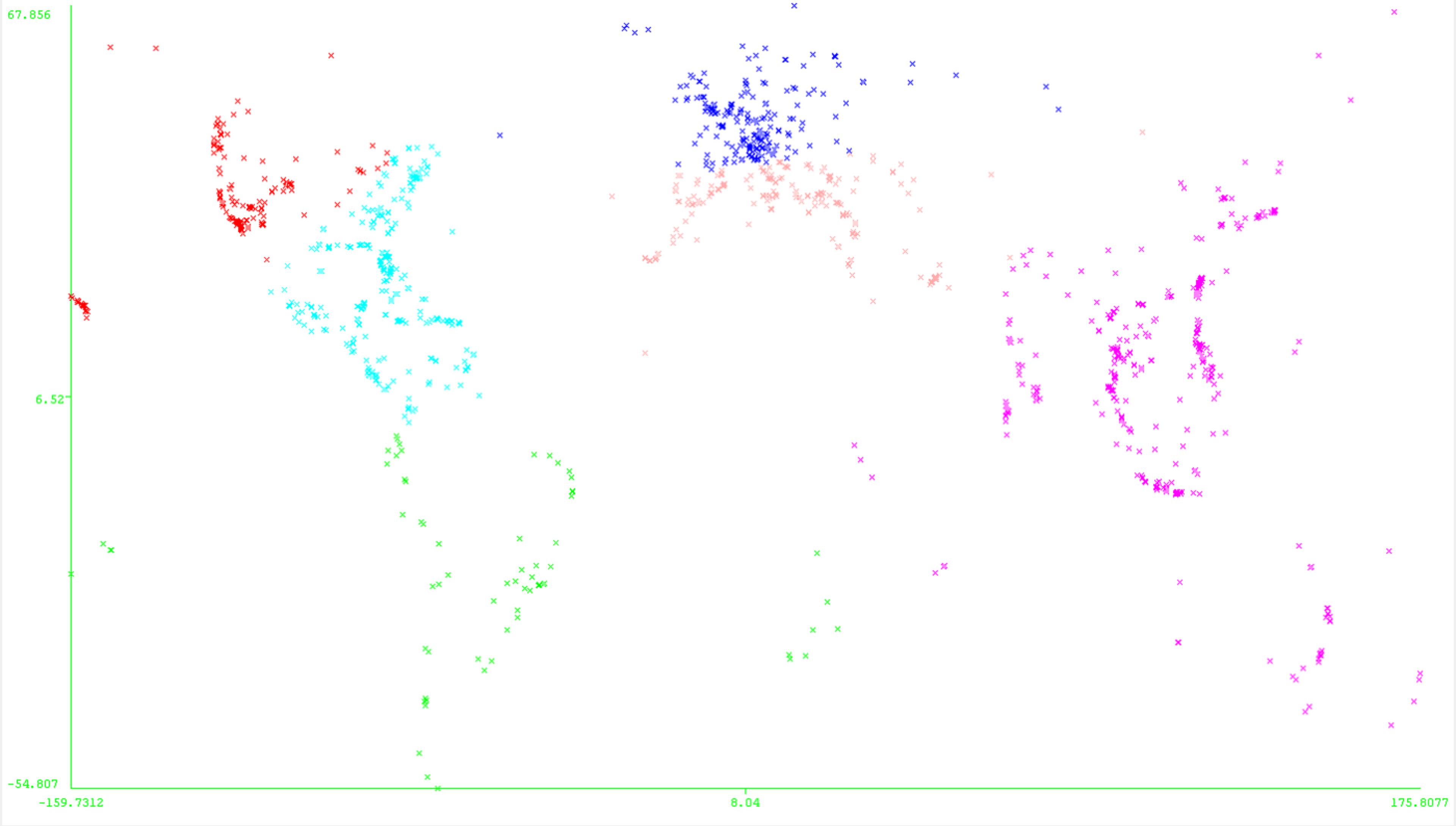
4



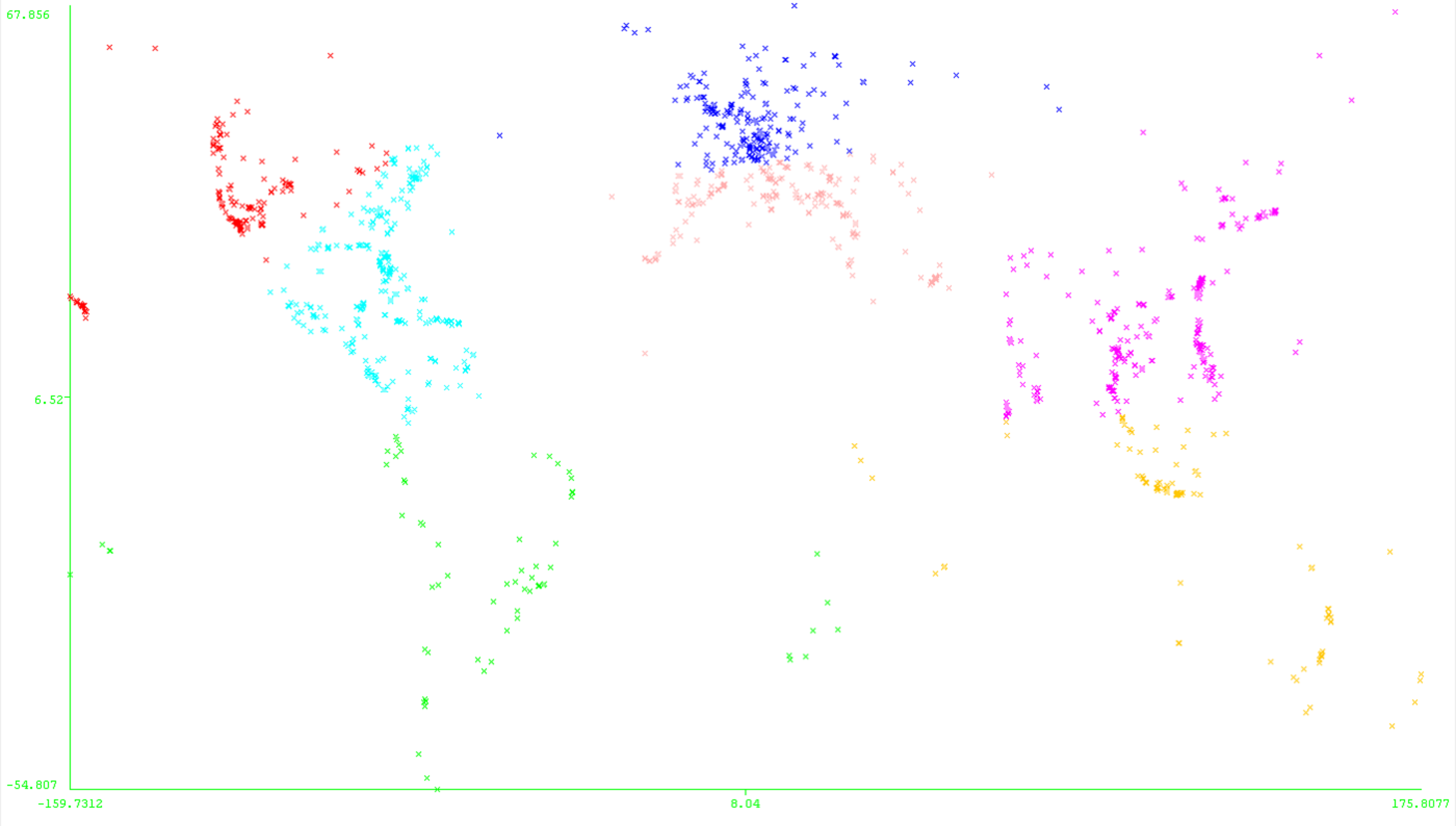
5



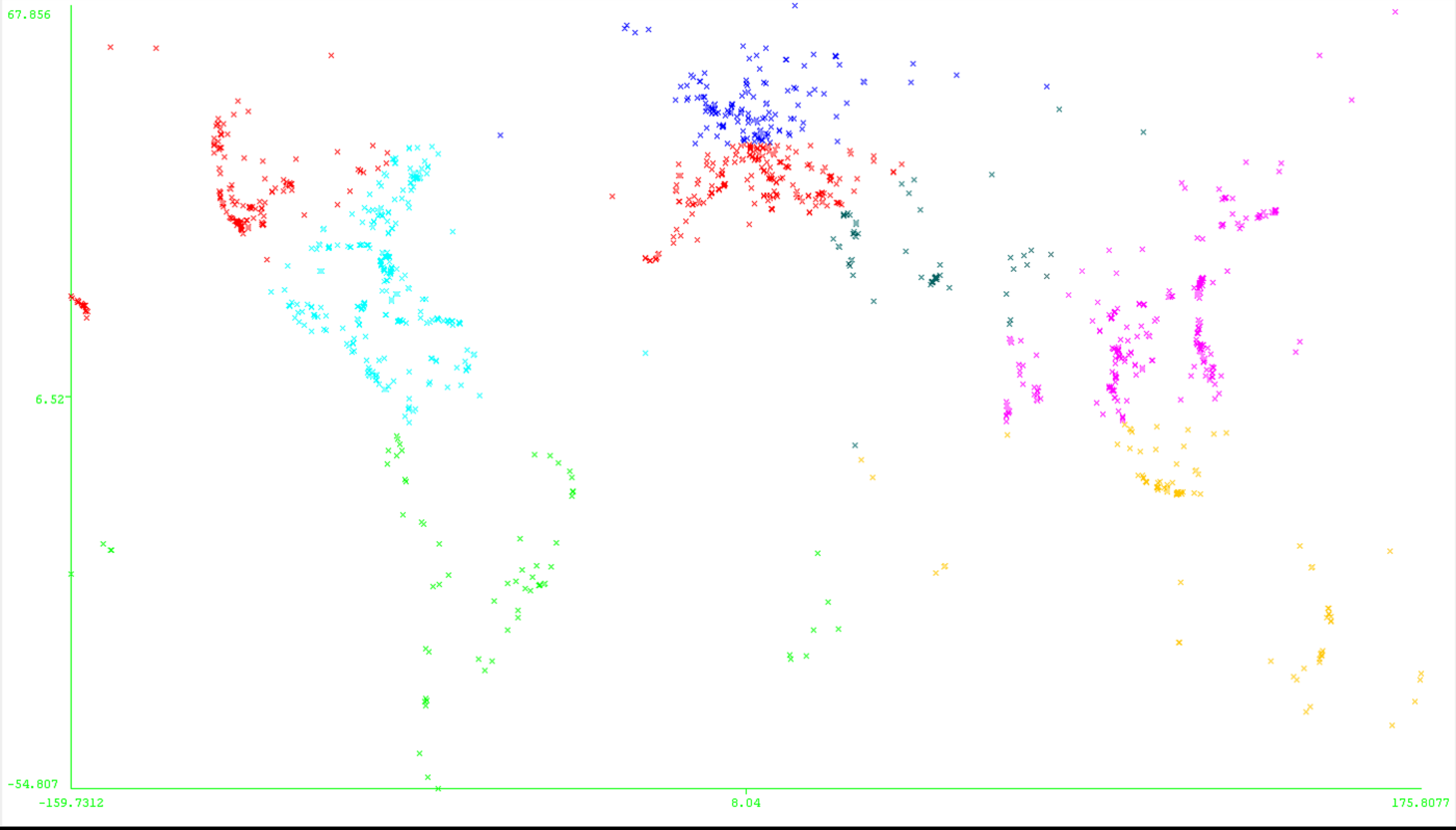
6



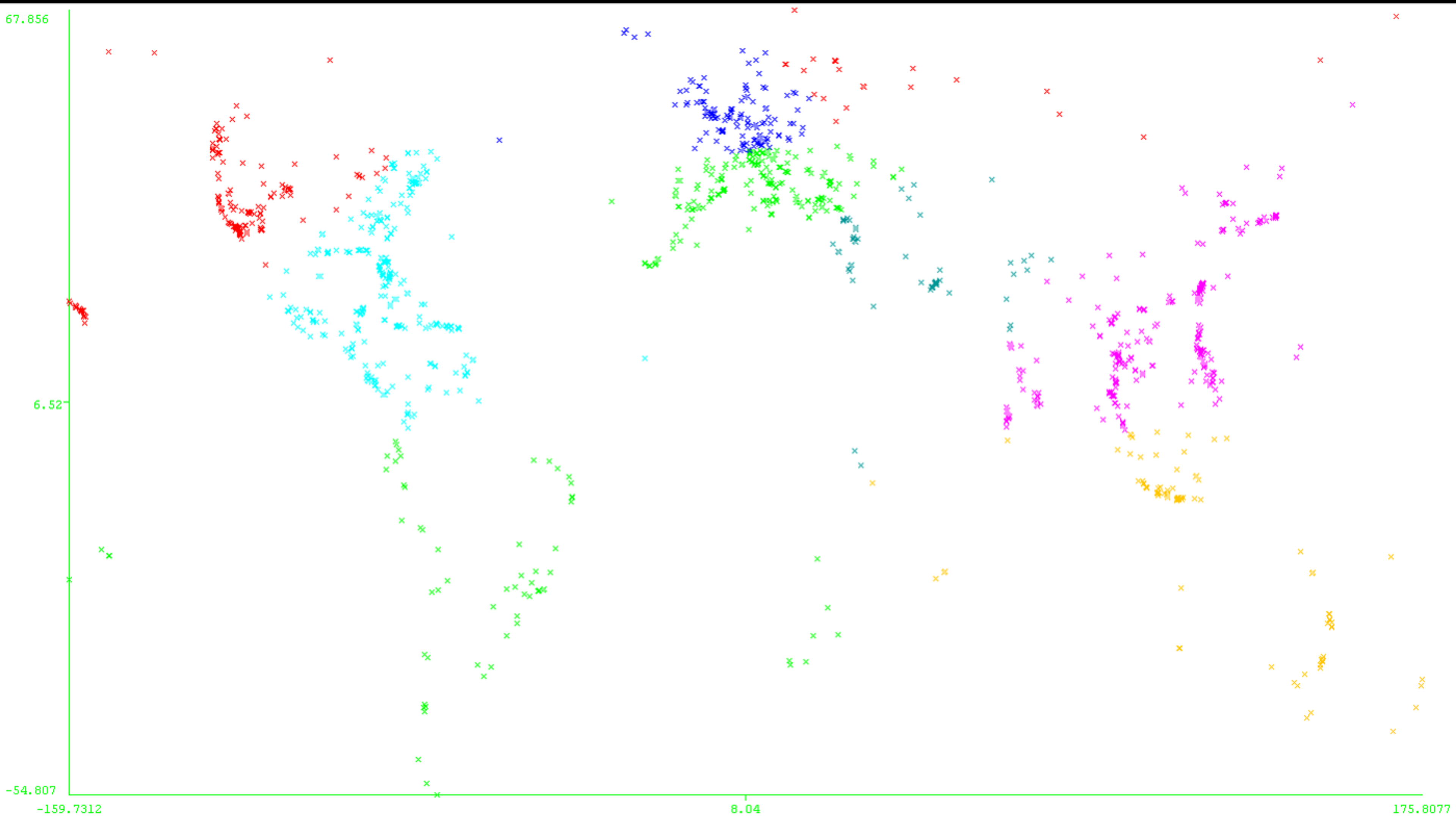
7



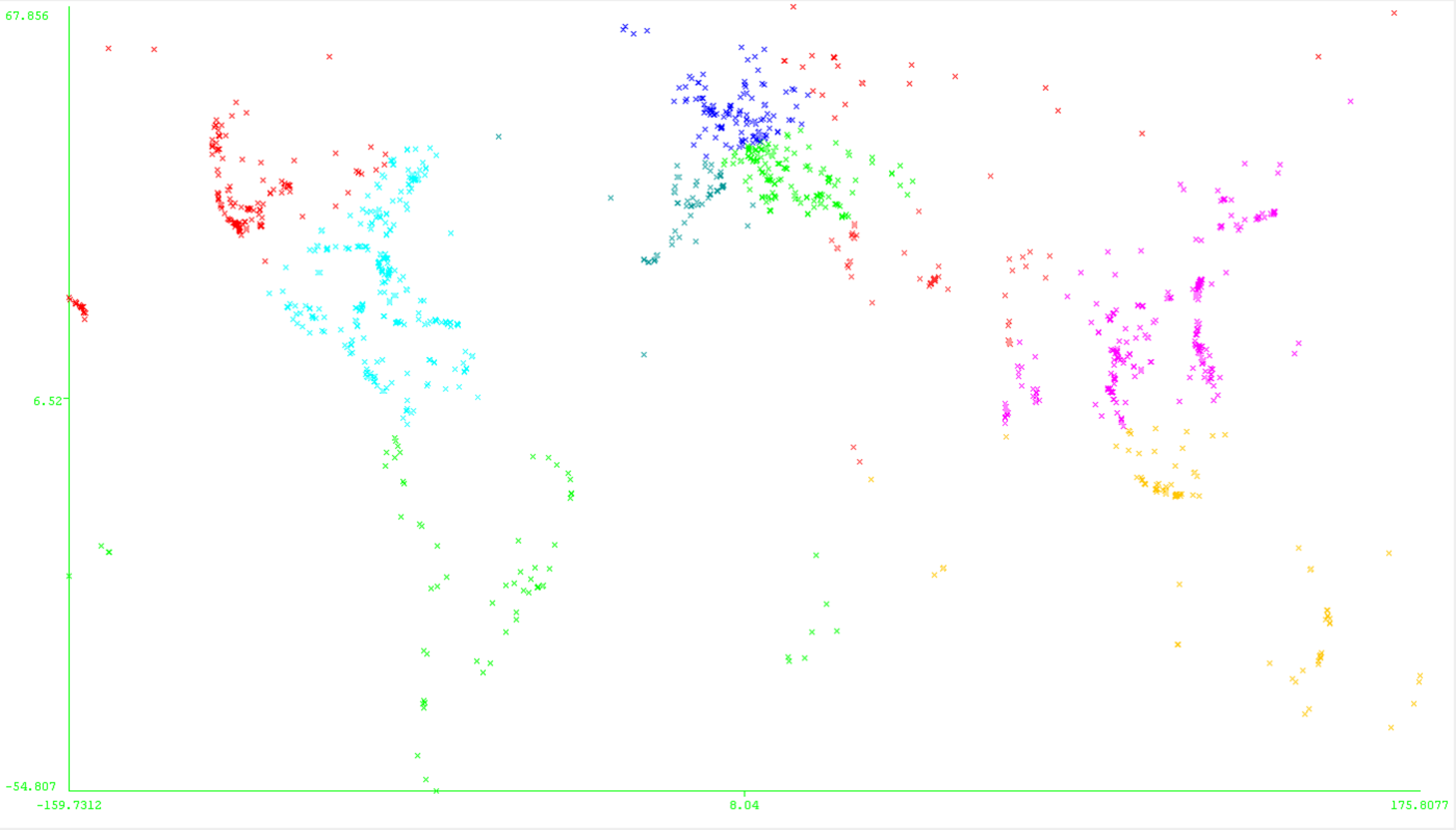
8



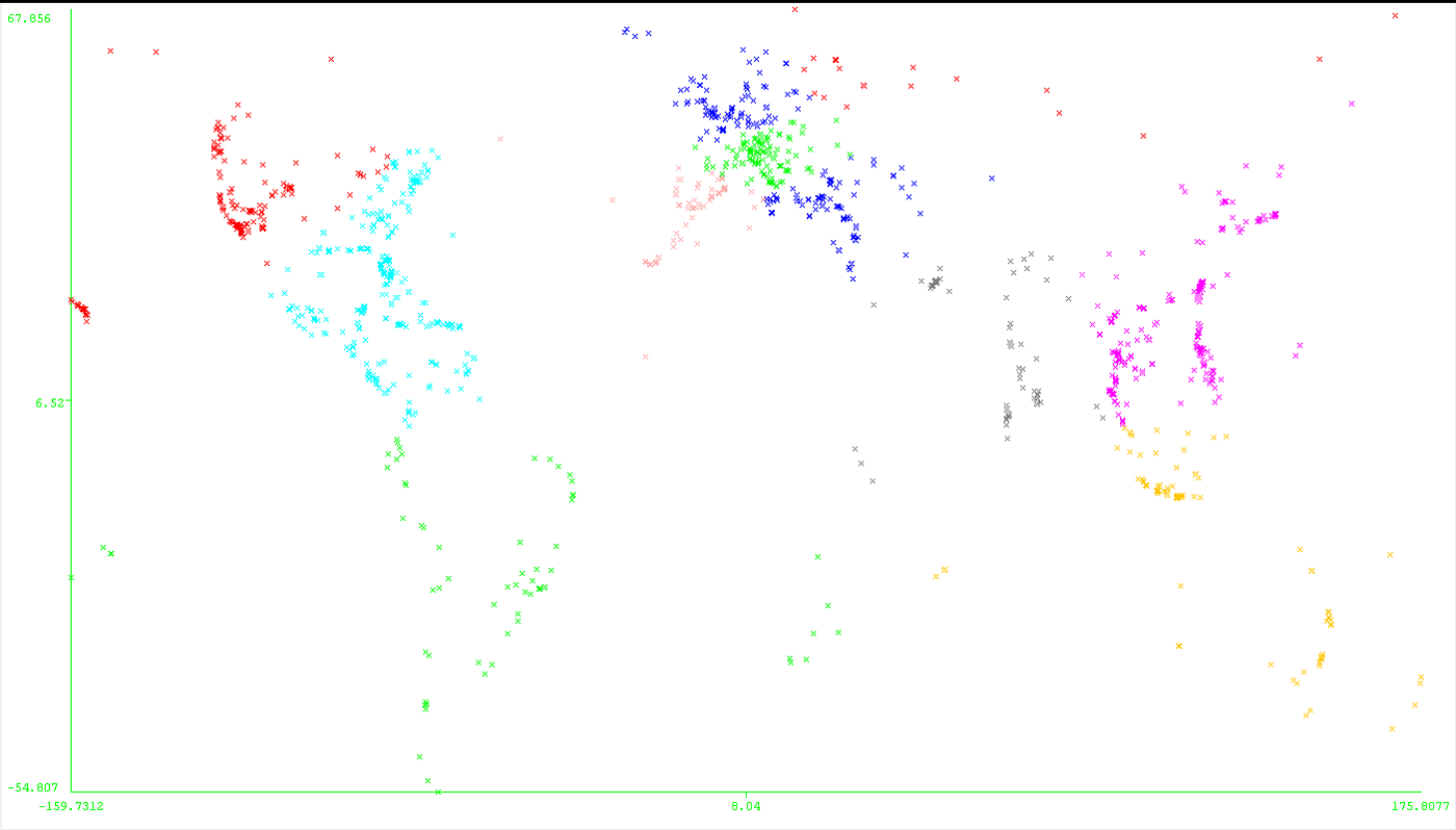
9



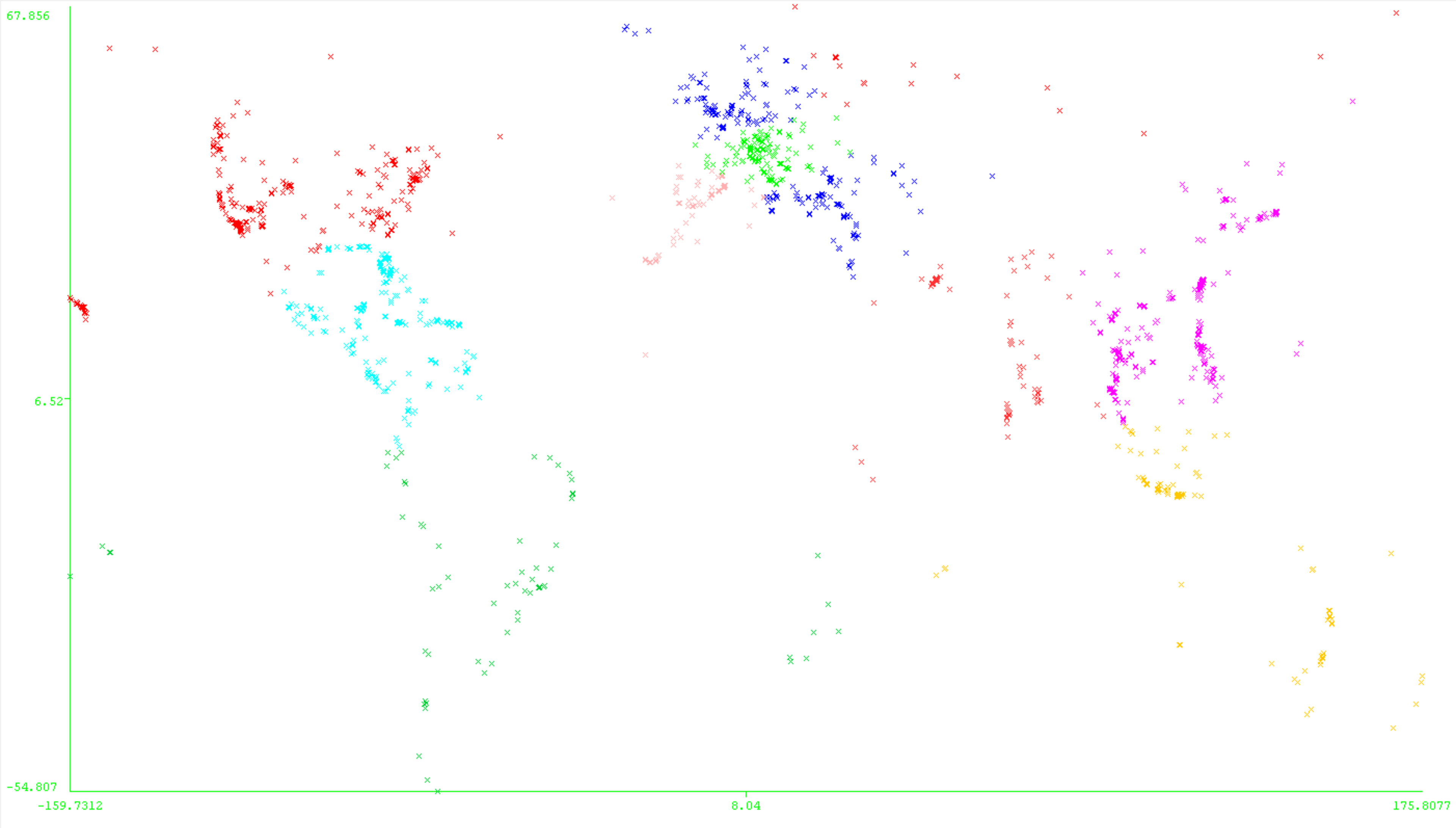
10



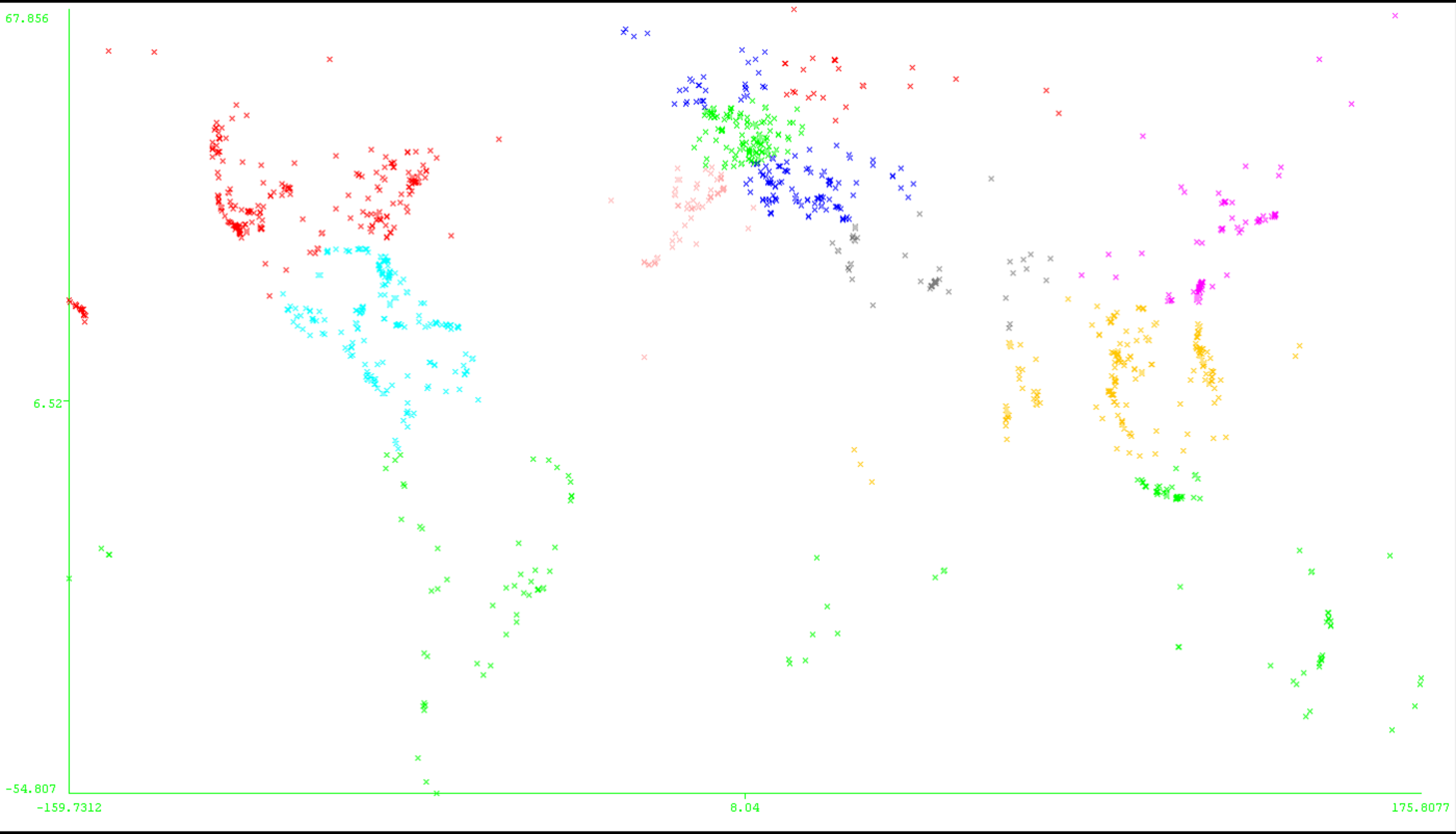
11



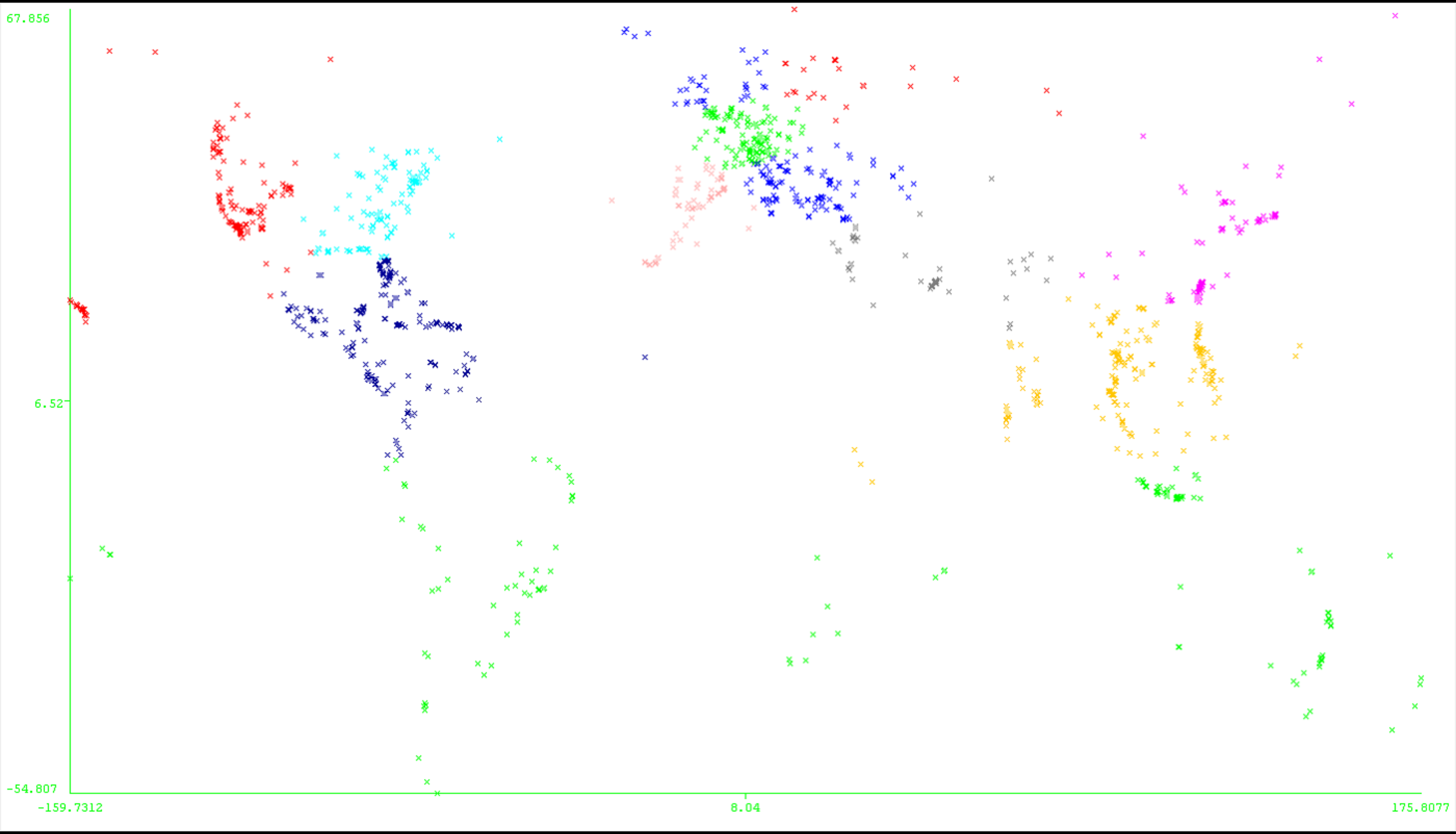
12



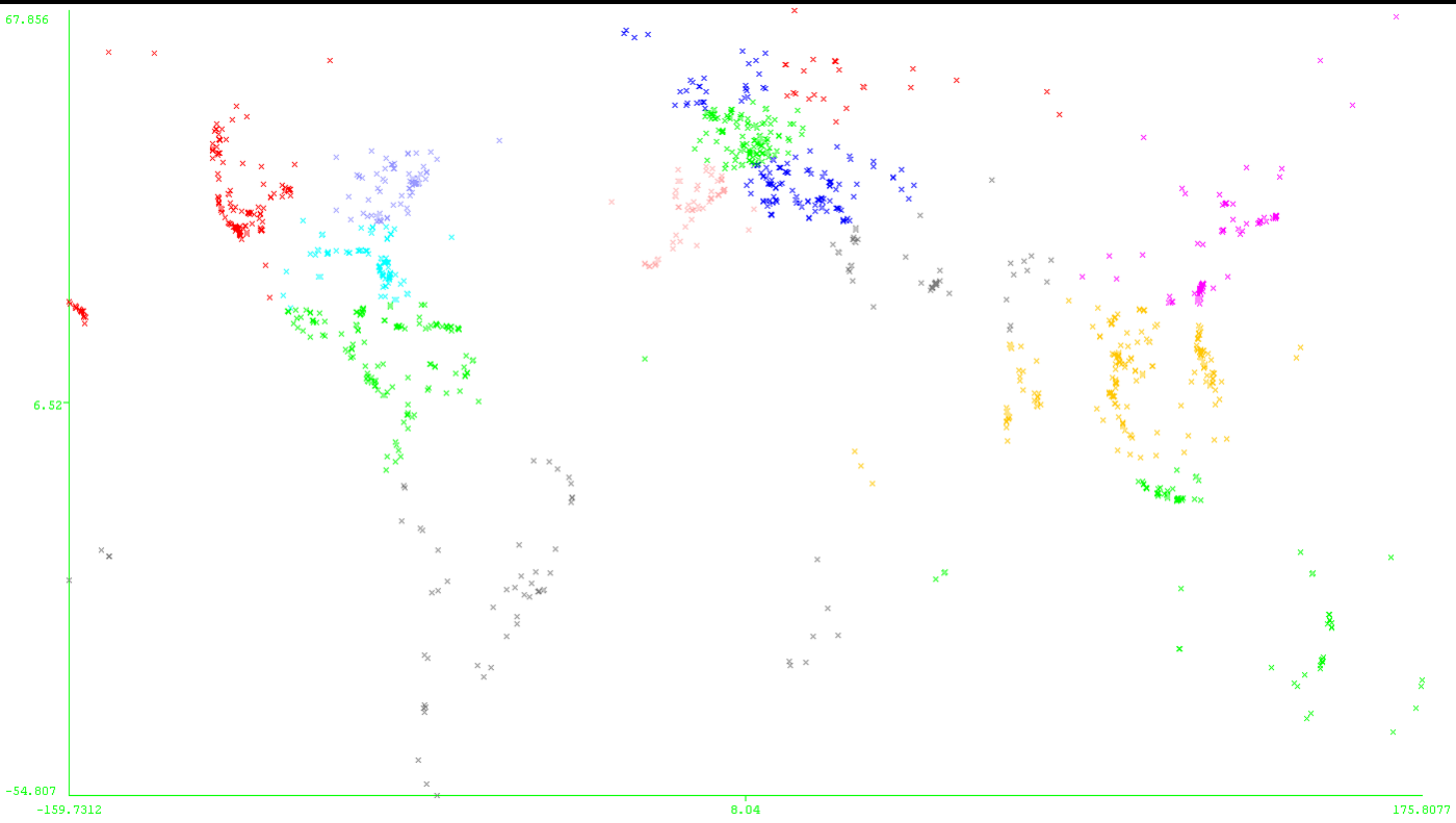
13



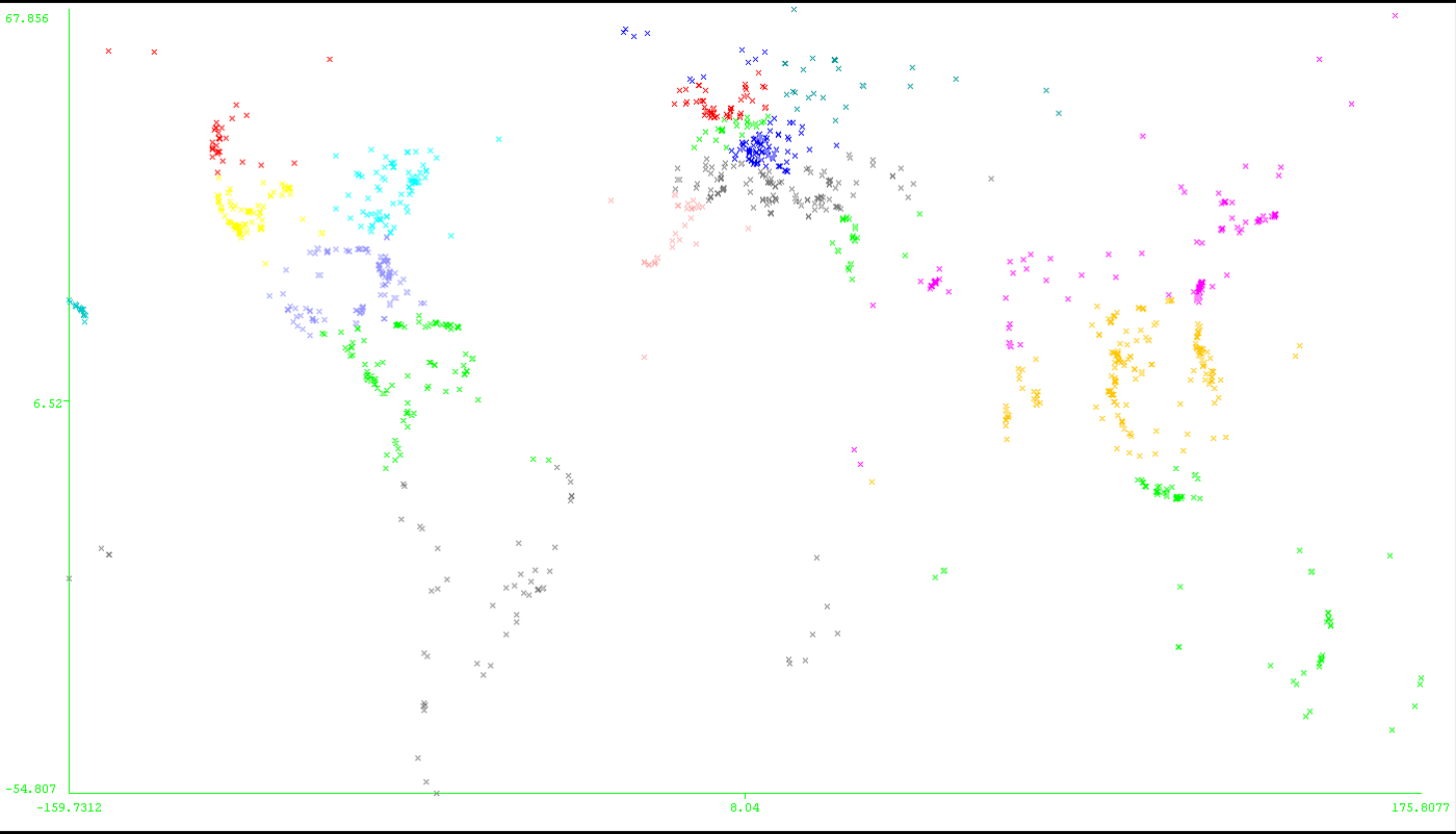
14



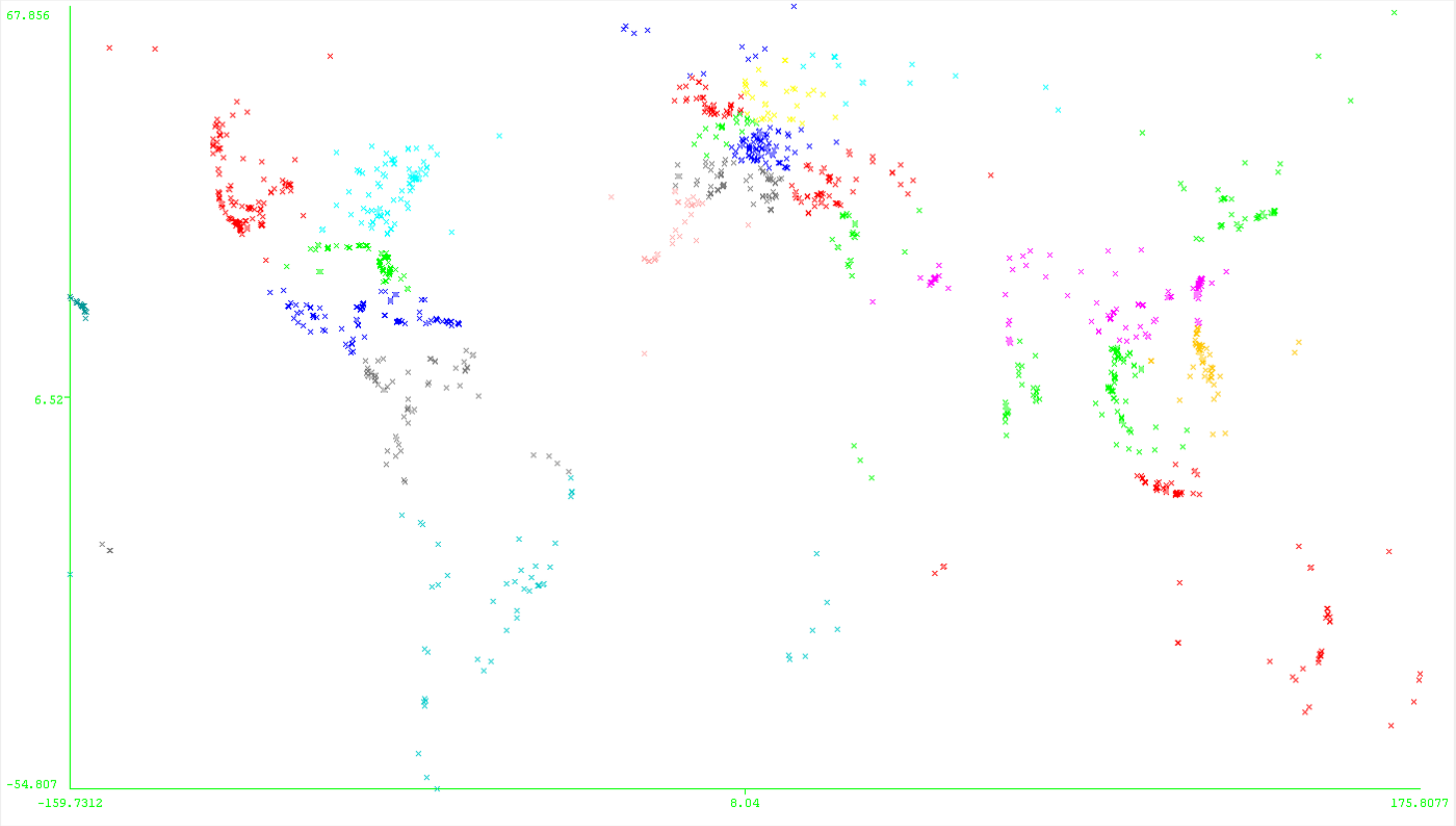
15



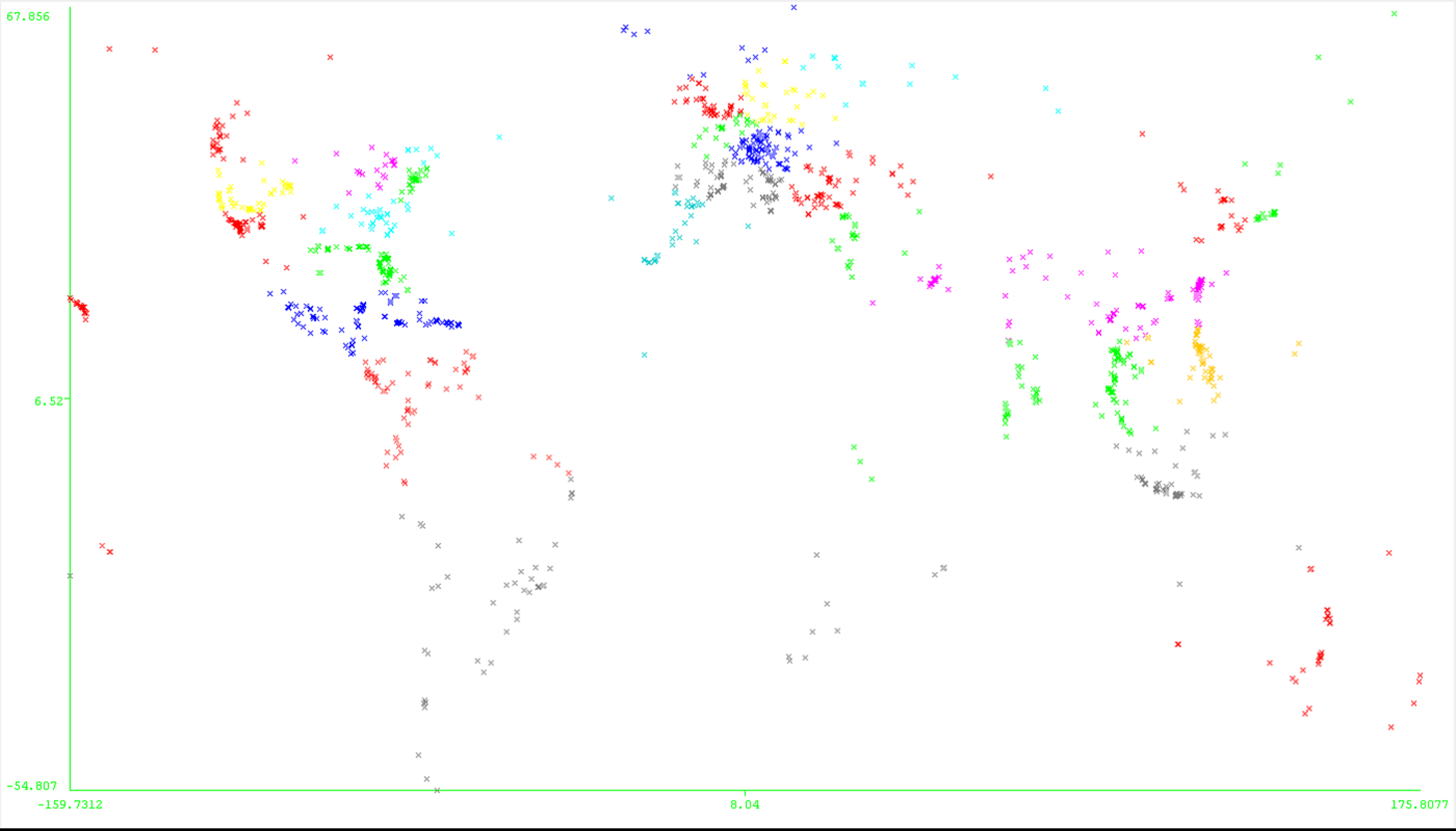
20



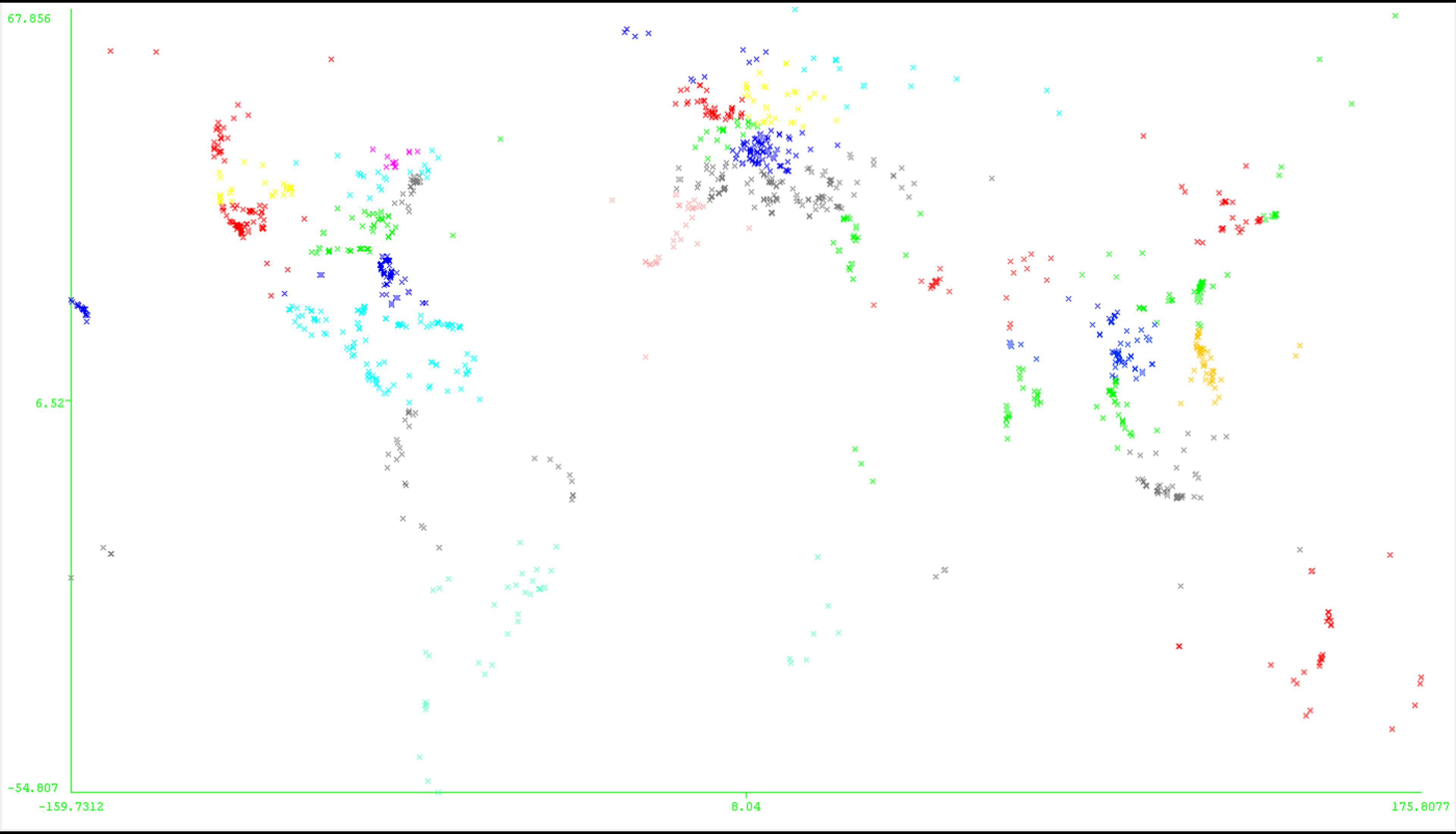
25



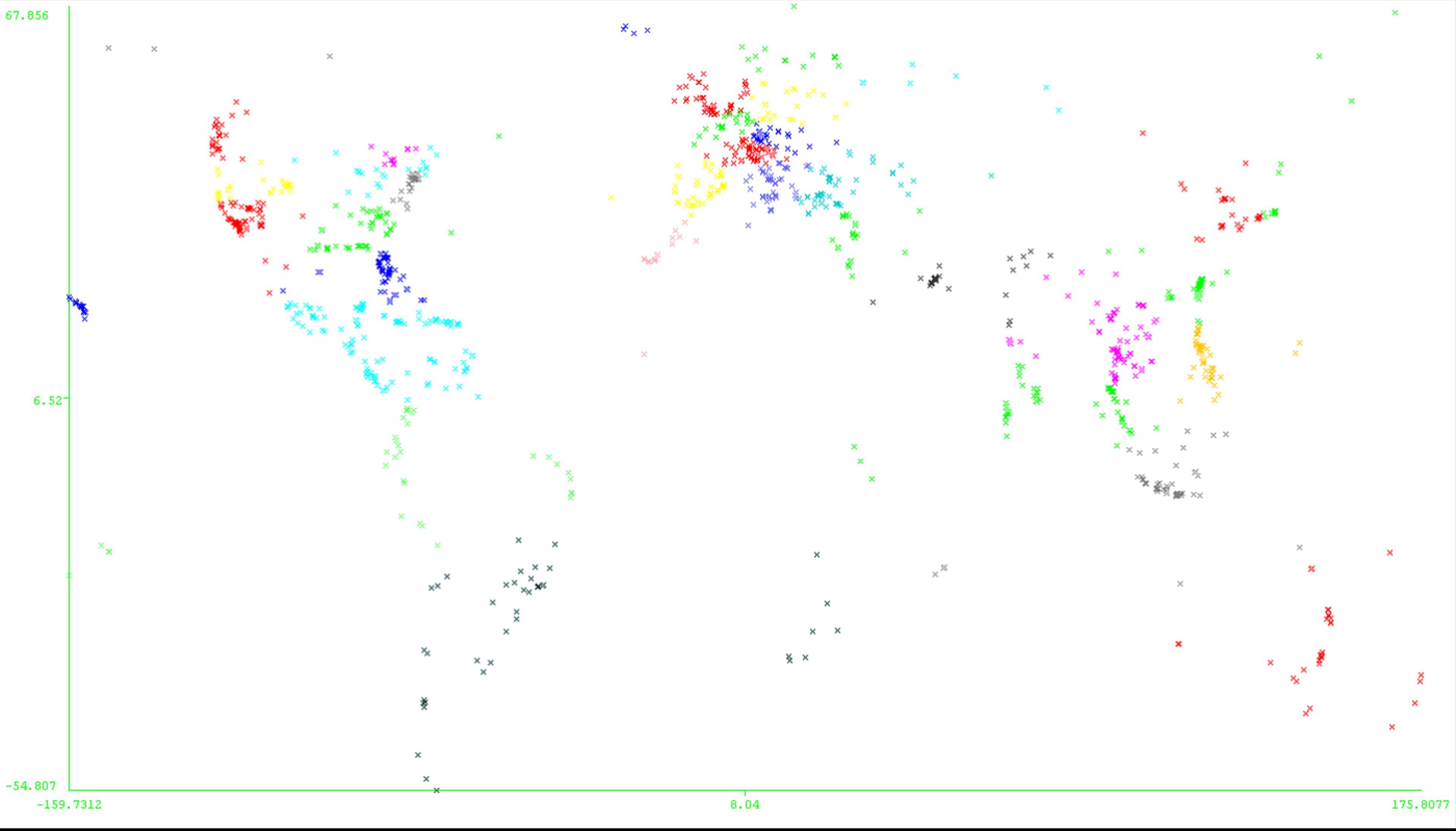
30



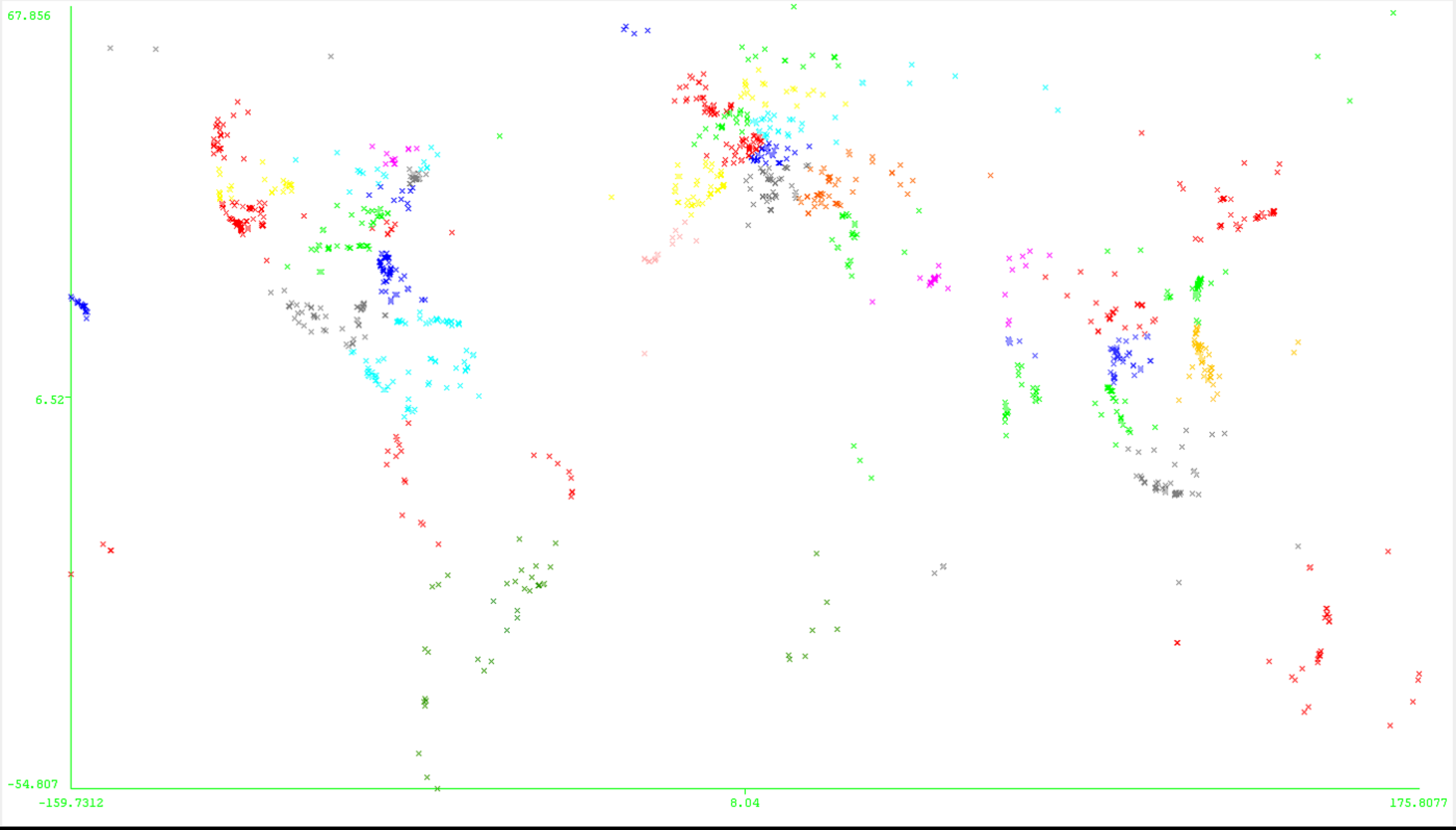
35



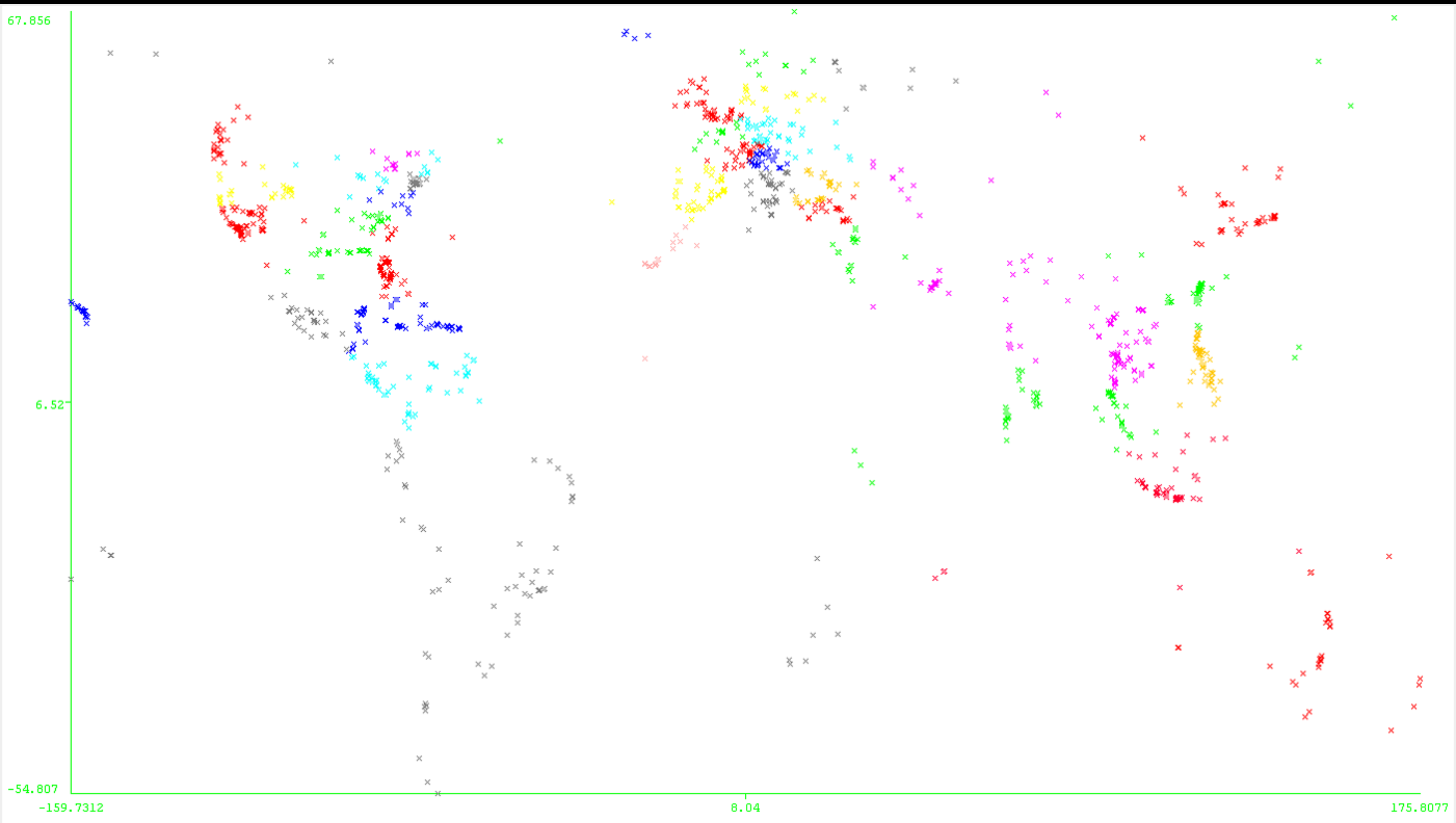
40

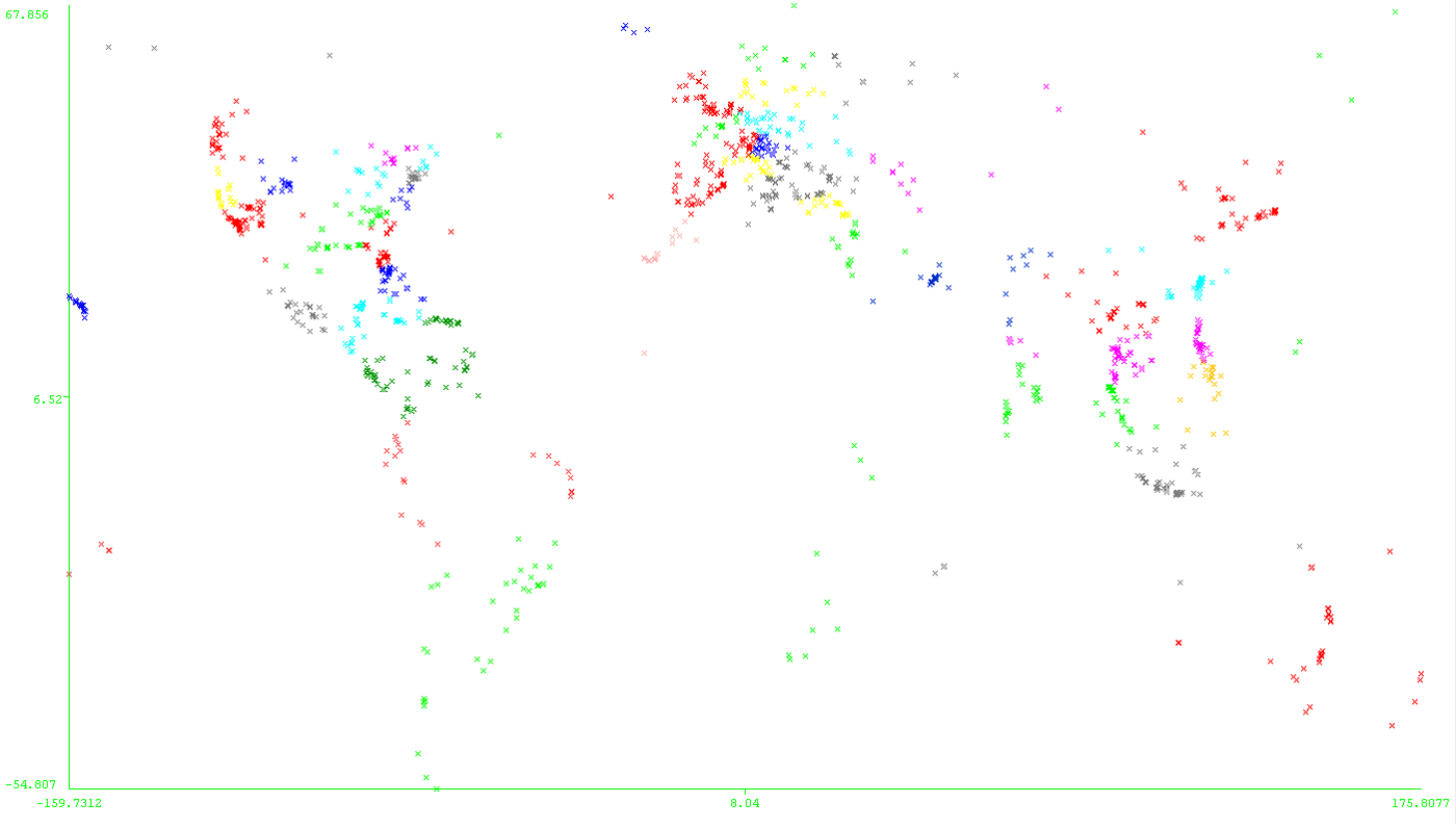


45

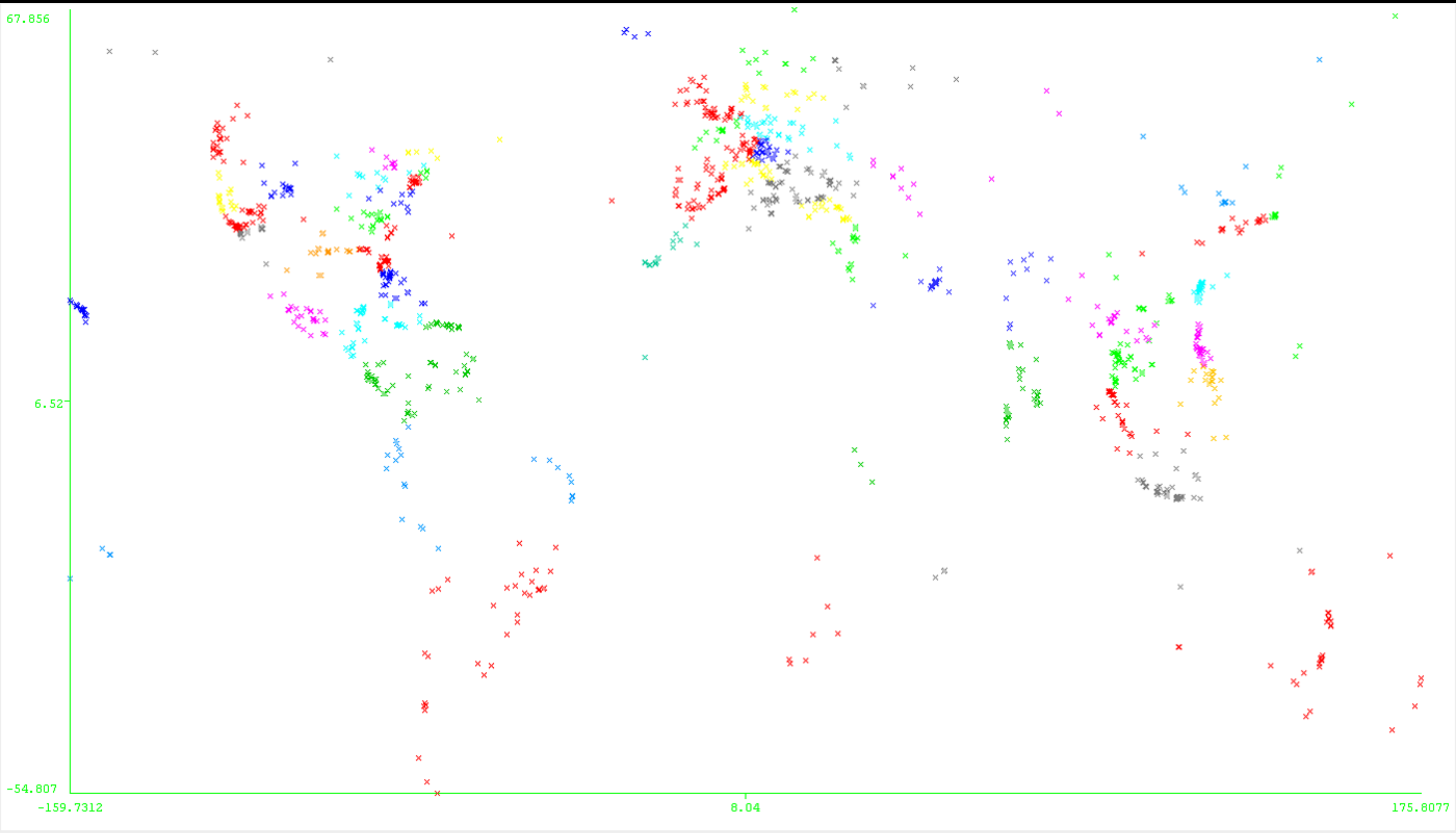


50

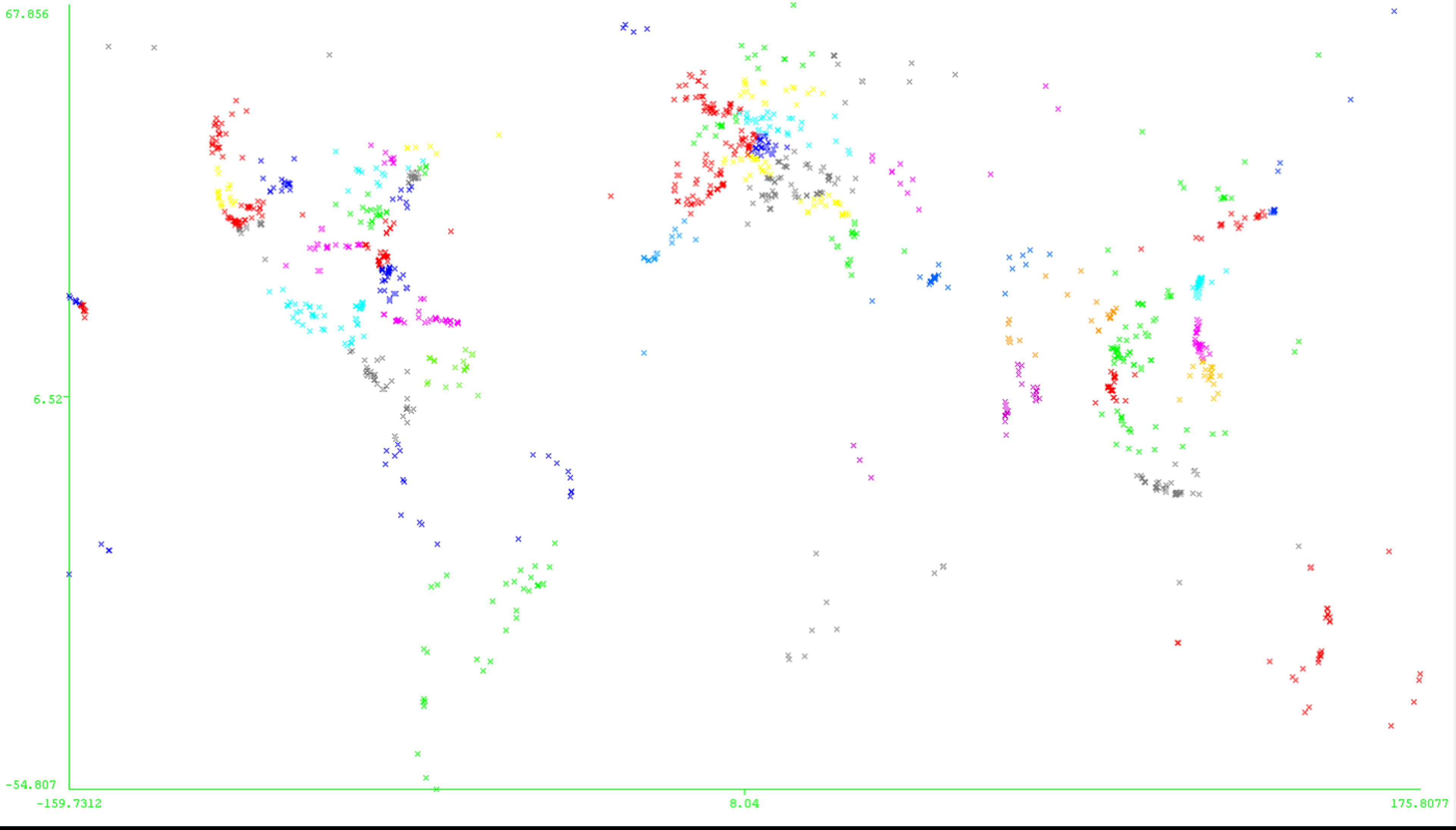



67.856
6.52
-54.807
-159.7312
8.04
175.8077

60



65



Results: Expectation Maximization

Cluster No.	No. of Points	% of Points	Mean: Lat	Mean: Long	Approximate Geographic Location
10	487	19%	26.6928	-82.0619	Florida
20	306	12%	10.0377	99.0563	Thailand
9	244	9%	50.4643	5.5763	Belgium/Lux/Germany
21	217	8%	35.5897	-116.6615	Las Vegas
8	157	6%	44.4515	11.598	Italy
0	140	5%	39.9862	20.8455	Greece
19	115	4%	35.7832	133.7258	Japan
3	109	4%	39.7927	-1.5861	Spain
16	94	4%	-21.0376	-45.0723	Brazil, RDJ
5	86	3%	-8.4773	114.5994	Malaysia/Singapore/Indonesia

- 22 Clusters
- Log likelihood: -8.32495

Results: X Means

Cluster No.	No. of Points	% of Points	Centroid: Lat	Centroid: Long	Approximate Geographic Location
35	147	6%	50.702845	1.002316	London
16	139	5%	45.432460	10.487077	Switzerland
0	111	4%	33.701722	-117.272512	Los Angeles
43	105	4%	39.765604	-1.848314	Spain
32	104	4%	25.738371	-80.248619	Miami
41	102	4%	41.787427	-75.853697	New York
49	75	3%	-8.477046	115.505835	Malaysia/Singapore/Indonesia
22	68	3%	41.786186	14.423605	Italy
24	66	3%	39.182161	30.409689	Turkey

- 68 Clusters
- Distortion: 38.612509

Results: K Means

Cluster No.	No. of Points	% of Points	Centroid: Lat	Centroid: Long	Approximate Geographic Location
20	112	4%	25.4566	-80.1601	Miami/Bahamas
29	112	4%	-8.398	114.4946	Malaysia/Singapore/Indonesia
51	105	4%	39.7656	-1.8483	Spain
7	103	4%	52.2395	0.0623	England
63	102	4%	19.8554	-95.5074	Mexico
33	84	3%	18.571	-70.1547	Caribbean
36	75	3%	42.7637	11.072	Italy
12	70	3%	13.7318	102.1059	Bangkok/Pattaya, Thailand
47	69	3%	28.3839	-82.013	Tampa/Orlando, Florida

- 65 Clusters
- Sum of Squared Errors: 2.027838695266866

Insights

- Euclidean distance for points in a 2D space
- Realize that data will have noise, decide whether to take it into consideration
- On simple data, most clustering algorithms perform fairly similar
- Investigate why Expectation Maximization works/ does not work.

Future Work

- I will expand (sugar coat) and sell this project to Thomas Cook
- Work with time series data
- Discover “Migration” patterns
- Use more than coordinates to determine insights about the data:
    1. Developing / Developed Locations
    2. Favorite Activities
    3. Idea time to travel
    4. Use more attributes #likes, #shares,

References

- https://apigee.com/console/instagram Testing the pull query
- https://netlytic.org/ pulling the dataset
- https://cran.r-project.org/web/packages analytics in R
    - DBSCAN
    - fcp
    - ggplot2, ggmap
    - stats
- http://ggplot2.org/ map plot in R
- Discovering latent clusters from geo-tagged beach images. Yang Wang and Liangliang Cao
- http://firstmonday.org/article/view/5563/4195, A methodology for mapping Instagram hash tags, Tim Highfield, Tama Leaver.
- http://www.ifp.illinois.edu/~cao4/papers/cao_icassp10.pdf, A worldwide tourism recommendation system based on Geotagged web photos, Liangliang Cao et. Al.