Movement-Based User Identity

Ilya Perepelitsa

Content

1. Data Transformation
2. Movement concept

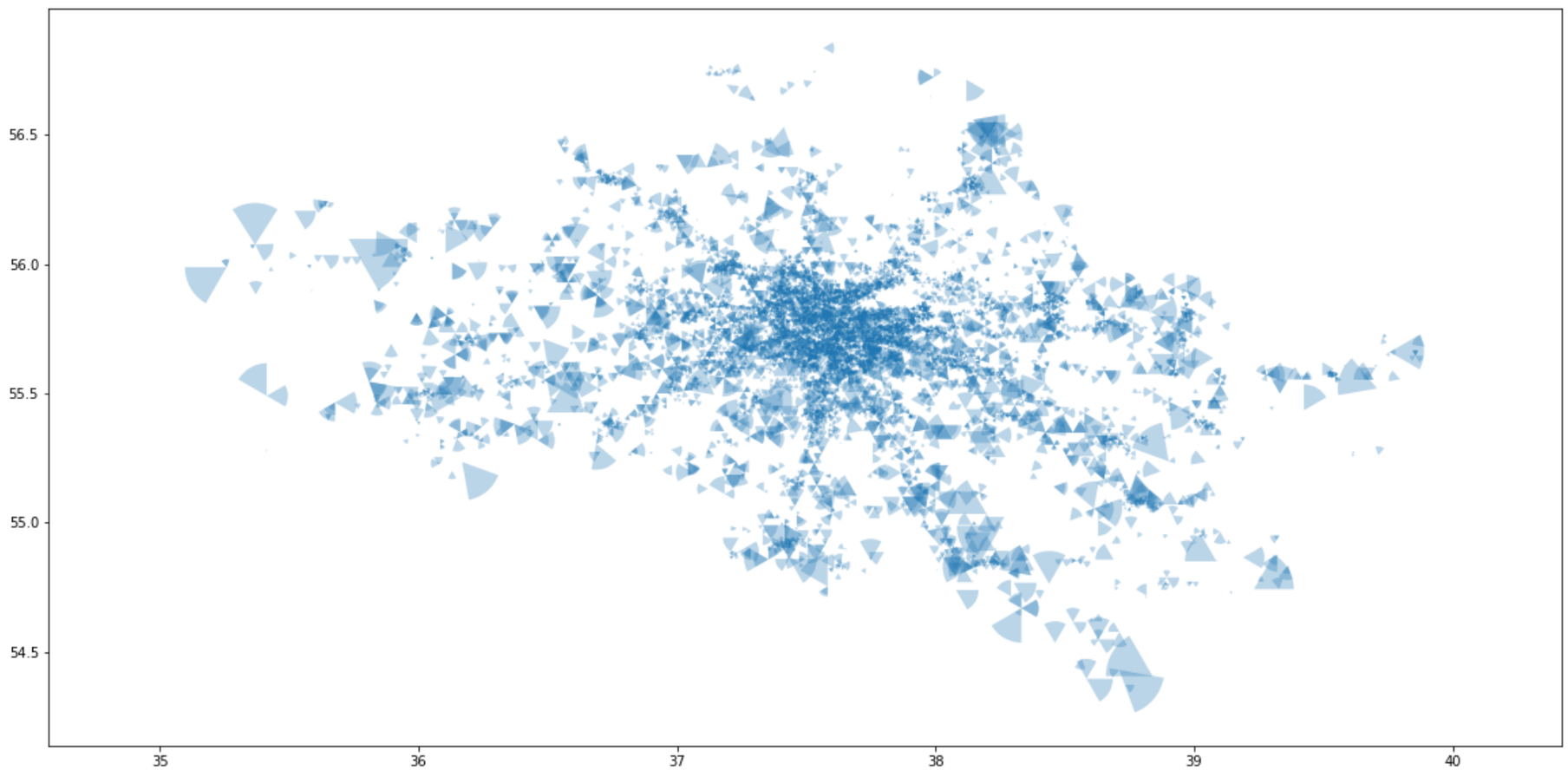
Data Transformation

Available location data

1. Tower location
2. Tower signal - angular data
3. Tower signal - radius

Meaning

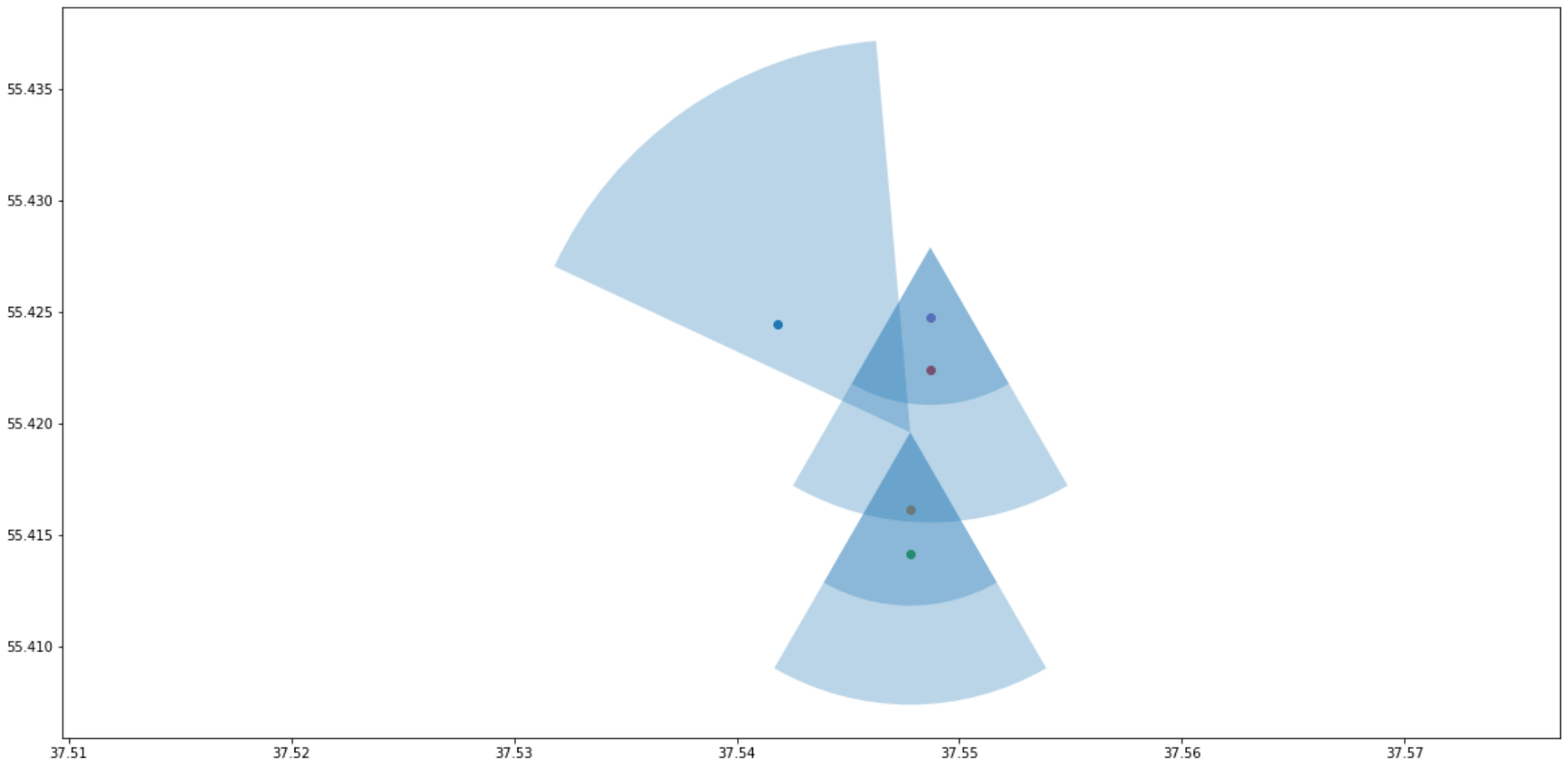
1. User location is only registered when an event occurs within the tower range
2. Other user movement isn't registered
3. Movement is "jumps between towers"
4. Distance between towers is regular travel distance



Data Transformation

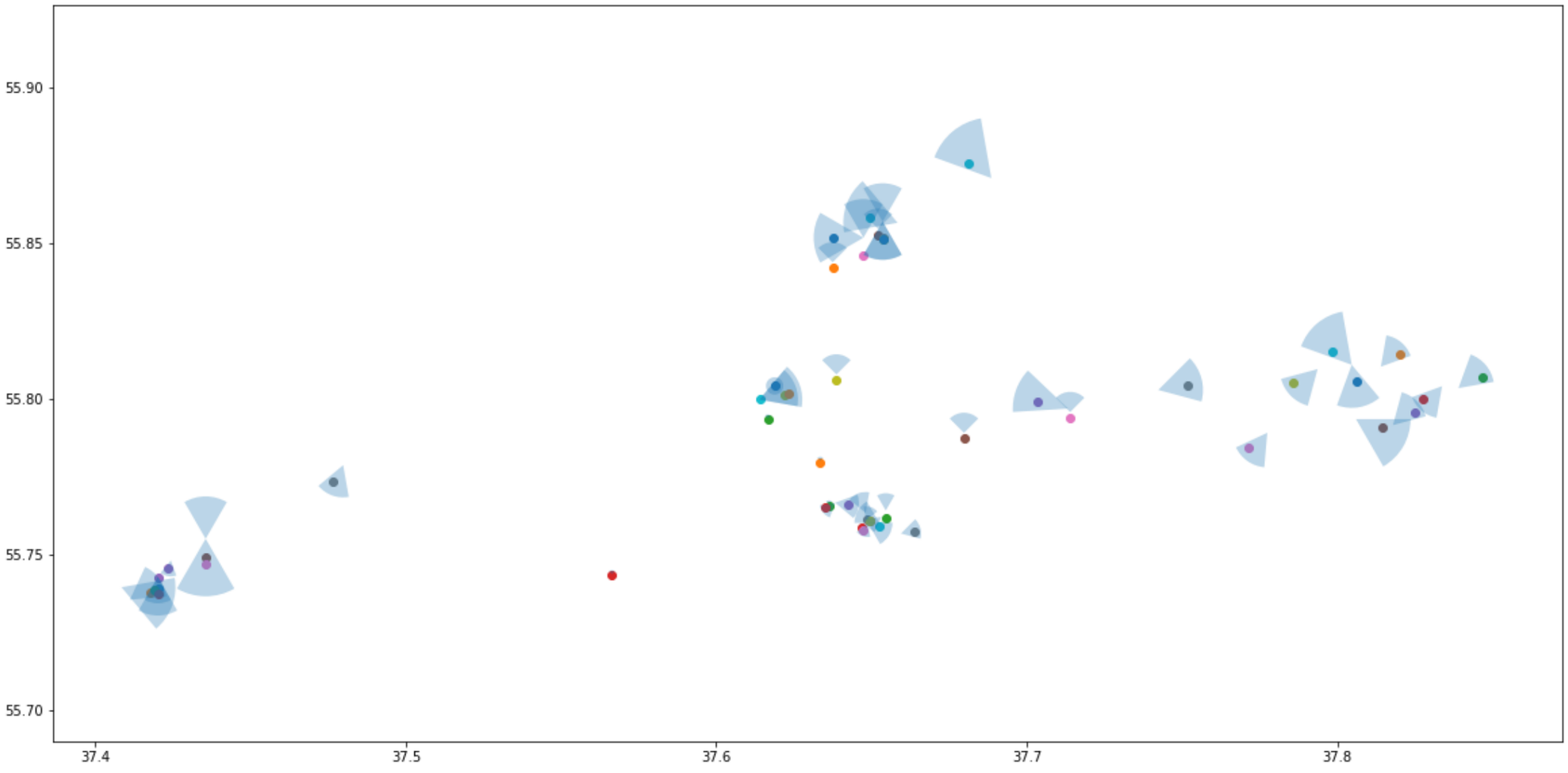
Problem	Solution

1. Tower locations don’t equal precise user location
2. Two towers with same location and different radiuses may register the same user at different time -> user moved
	1. Use circle sector centroid as a shortcut for location approximation
	  1. Reflects radius differences for the same location towers
	  2. Reflects what way the towers are facing



Data Transformation

Problem	Solution
1. Centroids misaligned	1. Ignore at this stage and troubleshoot later 2. Still better than raw location



Movement concept

If the pair of phone numbers is the same:

1) User uses them at the same time
2) User uses one device after using the other

In all cases

1) Locations mix
2) Locations “continue each other”

Conclusions:

1) If sorted by time the distance between updates shouldn’t be too large
2) Otherwise - long jumps between updates if plotted on the map as a line chart

msisdn	tstamp - **SORTED**	sector_centroid_lat	sector_centroid_lon
158599944901	2013-05-25 19:59:47.975	55.5337996	36.375054
158599999863	2013-05-25 19:08:50.816	55.65666	37.748562
158599999863	2013-05-25 18:14:57.490	55.65666	37.748562
158599999863	2013-05-25 17:17:46.001	55.65147	37.7413476
158599999863	2013-05-25 16:57:20.039	55.65147	37.7413476

Movement concept

Use:

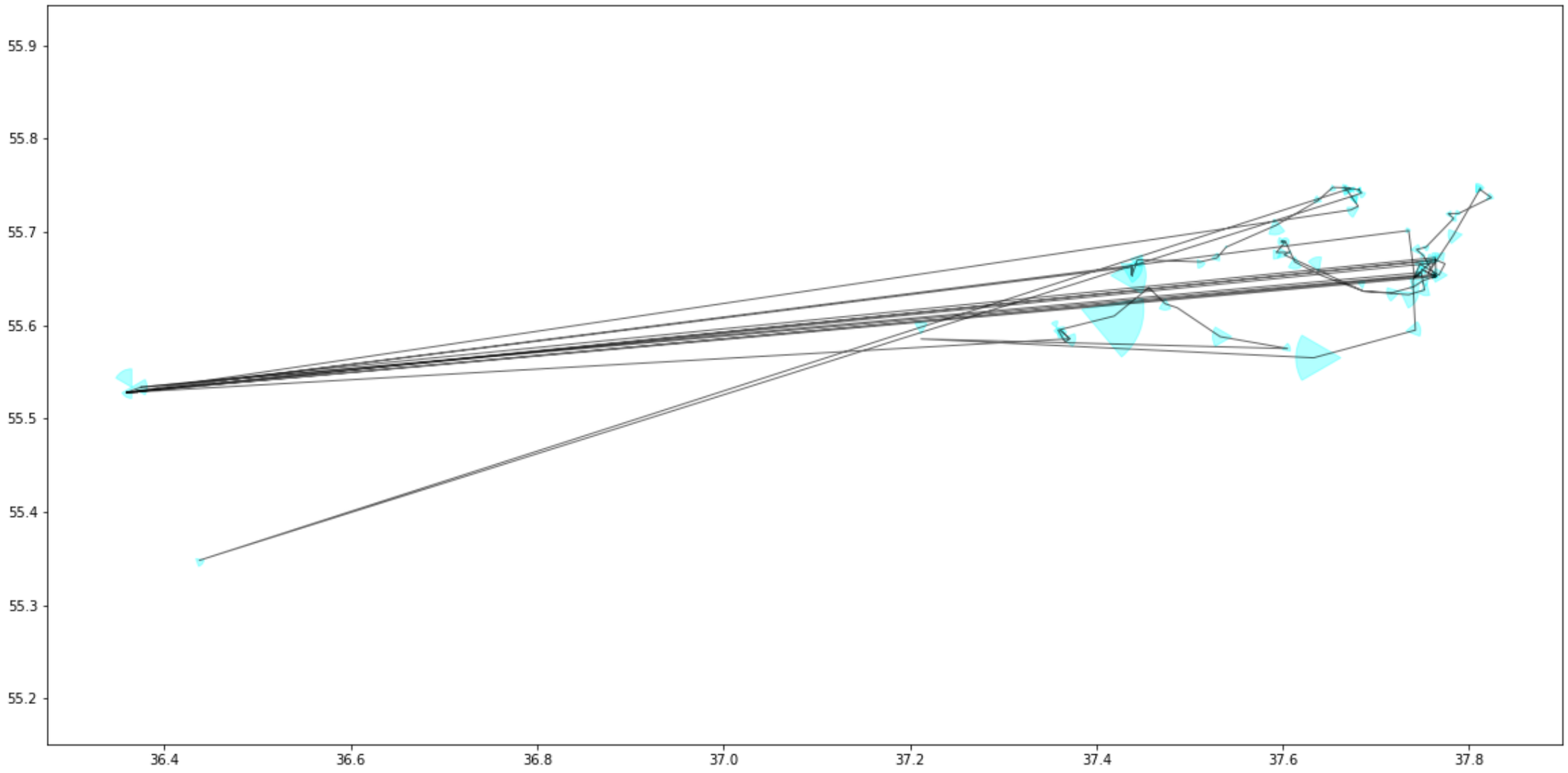
1) Observations where previous update isn't from the same phone number
2) Calculate time lag between such updates
3) Calculate geodesic distance between centroids
4) Calculate speed - distance travelled over time

Algorithm:

1) If the speed (average daily speed) is too great - the user couldn't have been the same, i.e. the updates are too far apart for them to belong to the same user

msisdn	msisdn_lag	tstamp	tstamp_lag	sector_centroid_lat	sector_centroid_lat_lag	sector_centroid_lon	sector_centroid_lon_lag
158599944901	158599999863	2013-05-25 19:59:47.975	2013-05-25 19:08:50.816	55.5337996	55.65666	36.375054	37.748562
158599999863	158599944901	2013-05-25 16:03:31.386	2013-05-25 16:03:18.568	55.7016324	55.528066	37.7350921	36.3592638
158599944901	158599999863	2013-05-25 16:03:18.568	2013-05-25 15:56:28.658	55.528066	55.7237588	36.3592638	37.6730885
158599999863	158599944901	2013-05-25 11:05:35.670	2013-05-25 10:55:47.713	55.6626301	55.528066	37.4377434	36.3592638
158599944901	158599999863	2013-05-25 10:55:47.713	2013-05-25 10:23:15.826	55.528066	55.5856147	36.3592638	37.3715612

Movement concept - map demo



Movement concept - engineered features

msisdn	msisdn_lag	bbox_25p	bbox_50p	bbox_75p	bbox_mean
158599978993	158599999863	14545.93	32586.43	95394.66	47405.28
158599971737	158599999863	14545.93	32586.43	95394.66	47405.28
158599955935	158599999863	14545.93	32586.43	95394.66	47405.28
158599944901	158599999863	14545.93	32586.43	95394.66	47405.28
158599940677	158599999863	14545.93	32586.43	95394.66	47405.28

msisdn	msisdn_lag	speed_25p	speed_50p	speed_75p	speed_mean
158599978993	158599999863	0.00	2.40	27.71	132.24
158599971737	158599999863	3.75	26.52	71.73	305.13
158599955935	158599999863	2.65	20.64	53.60	218.99
158599944901	158599999863	2.25	18.33	56.84	529.09
158599940677	158599999863	3.53	23.15	72.29	347.91

Features:

1) Binding box diagonal - all aforementioned events are within a lat-lon-min -> lat_lon_max box, the diagonal is the distance between two points. Engineered features are the daily:
    1) 25 percentile of diagonal distance in meters
    2) 50 percentile
    3) 75 percentile
    4) mean
2) Speed - the distance traveled divided by the time delta between two events, kilometers per hour
    1) 25 percentile of speed
    2) 50 percentile
    3) 75 percentile
    4) mean

Problems:

1) 2.75 pairwise phone number combinations - time is needed to perform next steps