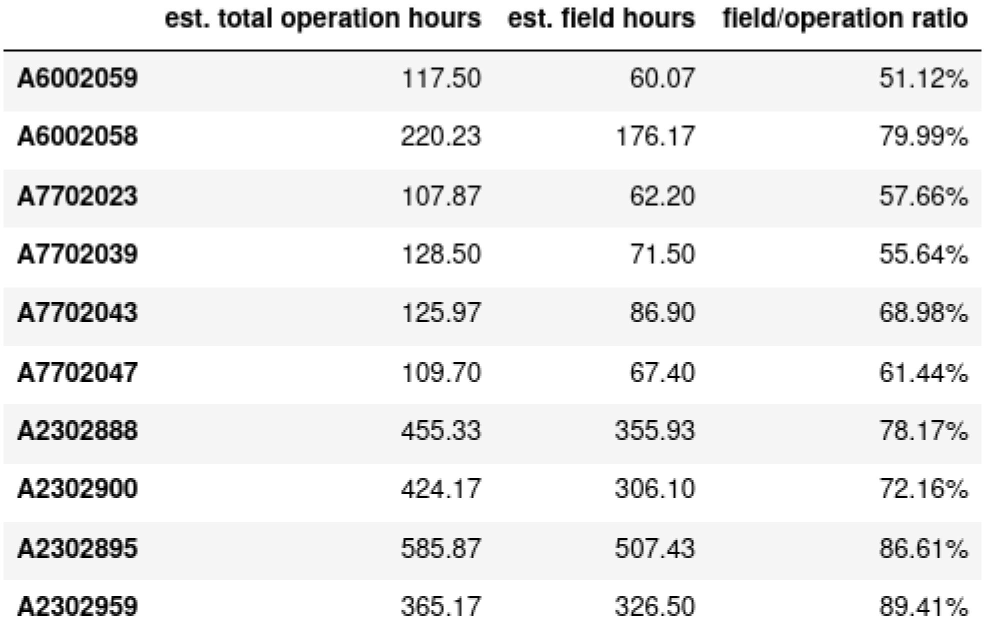
Analysis of farming mechanization telemetry

- *Candidate: Aleksa Radosavčević (aleksaradosavcevic@gmail.com)*
- *20th of March 2023, Kragujevac, Serbia.*

1. What is the ratio between transport and actual field work?

In order to process data with time and compute resources at disposal, the time series for each machine were resampled to 8 minute interval observations, whereas longitude and latitude were aggregated by averaging and rounding to 6 decimals. Next, for each machine, time series were sorted by timestamp. The ratio between field work and transportation was approximated by geopy's *geolocator.reverse* functionality which, for given longitude and latitude, returns an address. Simply, If "road" information was present in the address dictionary, time series observation was marked as transportation, otherwise observation was marked as field work.

Observations where machines were in a standstill mode were removed, and in the last step, from the portion of the machines' working hours, the field work utilization percentage was calculated. The remaining percentage difference is attributed solely to transportation:



	est. total operation hours	est. field hours	field/operation ratio
A6002059	117.50	60.07	51.12%
A6002058	220.23	176.17	79.99%
A7702023	107.87	62.20	57.66%
A7702039	128.50	71.50	55.64%
A7702043	125.97	86.90	68.98%
A7702047	109.70	67.40	61.44%
A2302888	455.33	355.93	78.17%
A2302900	424.17	306.10	72.16%
A2302895	585.87	507.43	86.61%
A2302959	365.17	326.50	89.41%

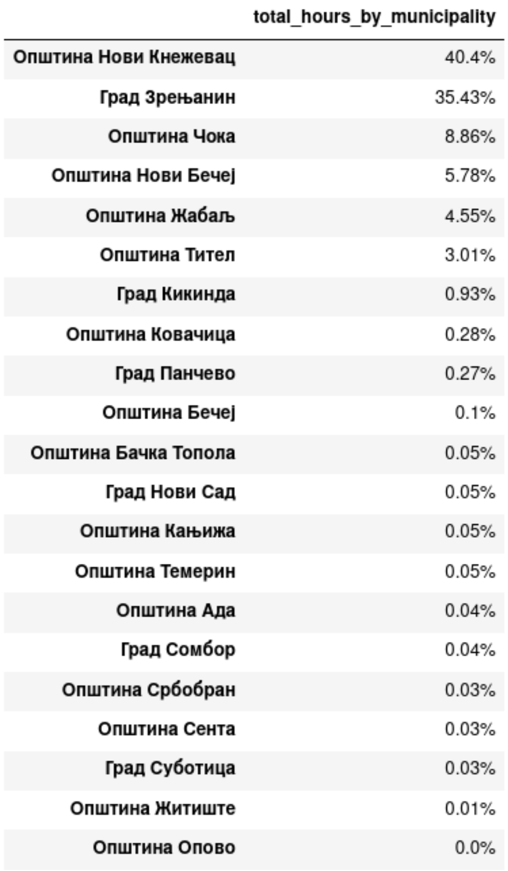
Dispersion in utilization percentages could be explained by different machine types (i.e. different by size, power…), different sizes of fields where machines have been working on, covering significantly larger areas. Moreover, by age - the older the machine is, the utilization is expected to be higher and vice-versa for newer machines - some time is needed to pass in order to increase the number of field hours. Lastly, by lower yields - if yields were lower than expected some machines could be under-utilized.

Alternative solution: If “road” information wouldn’t be present at all, a possible way to distinguish transportation from field work would be by analyzing percentiles of each machine *speed* variable. Then, a simple heuristic approach could be to localize the longest consecutive speed peaks subsequences in time series, when the speed threshold is exceeded (with some break tolerance, i.e 1 or 2 observations could be lower - if there are some traffic light or intersections during transport), and marking those observations’ indices as transportation mode. This could be implemented via scipy.signal.

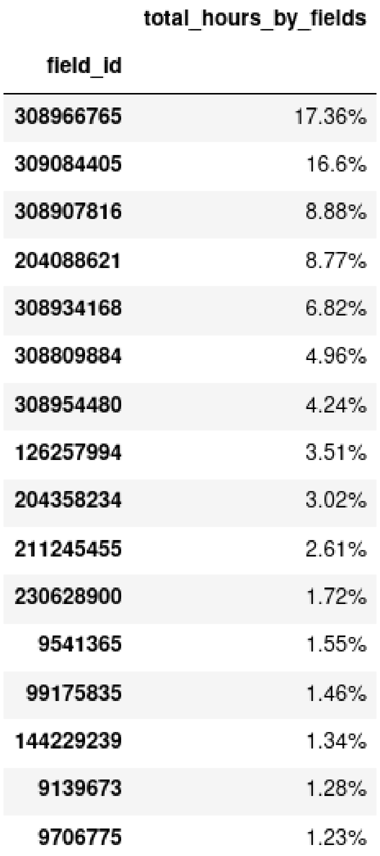
2. How many different fields (plots) did mechanization work on?

To obtain the number of fields processed by each machine, “place_id” of recognized “addresses” by geopy’s geolocator functionality were taken as a reference. After cross-comparison of fields’ IDs of different machines, overlapping field IDs among machines were noted, indicating that more machines were used to process single fields. This supports the aforementioned hypothesis regarding the existence of a set of larger individual fields or dense fields’ cluster(s).

By applying this method, 1584 unique fields were estimated (displayed only 16 field IDs with time allocation higher than 1%), allocated over 21 municipalities:

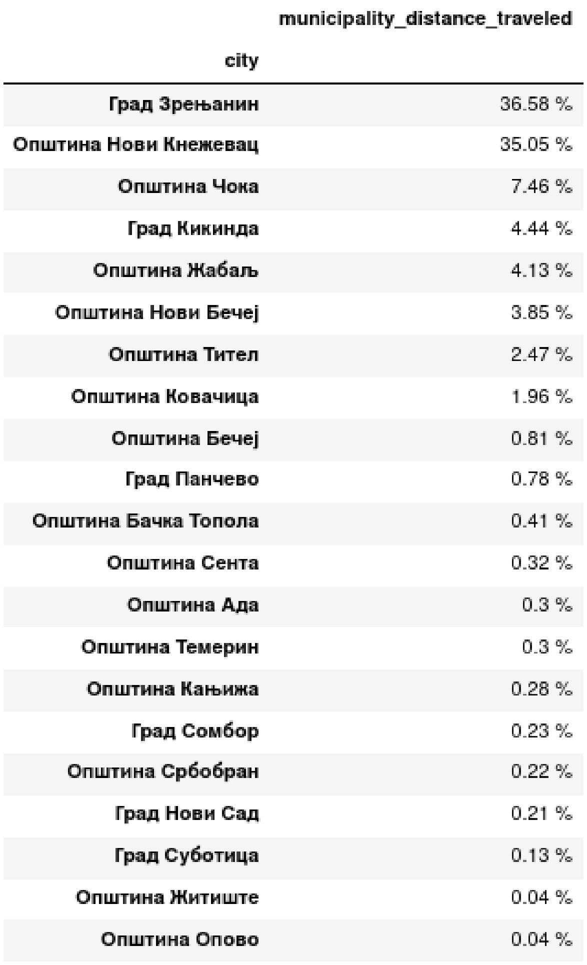


	total_hours_by_municipality
Општина Нови Кнежевац	40.4%
Град Зрењанин	35.43%
Општина Чока	8.86%
Општина Нови Бечеј	5.78%
Општина Жабаљ	4.55%
Општина Тител	3.01%
Град Кикинда	0.93%
Општина Ковачица	0.28%
Град Панчево	0.27%
Општина Бечеј	0.1%
Општина Бачка Топола	0.05%
Град Нови Сад	0.05%
Општина Кањижа	0.05%
Општина Темерин	0.05%
Општина Ада	0.04%
Град Сомбор	0.04%
Општина Србобран	0.03%
Општина Сента	0.03%
Град Суботица	0.03%
Општина Житиште	0.01%
Општина Опово	0.0%

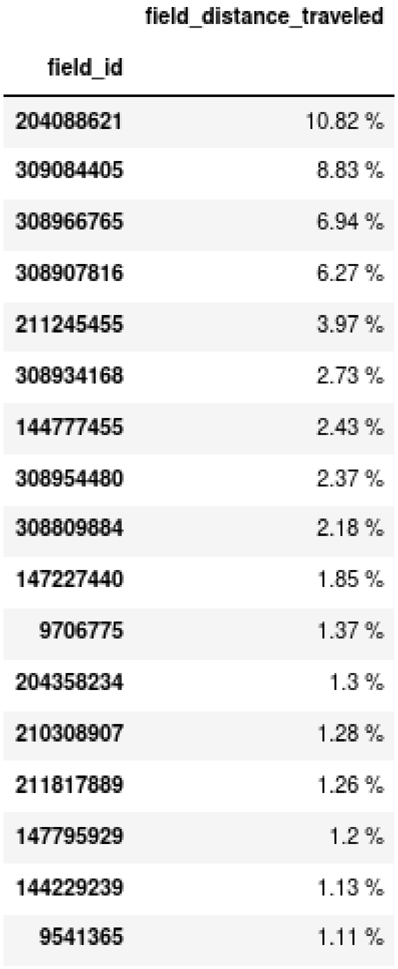


	total_hours_by_fields
field_id	
308966765	17.36%
309084405	16.6%
308907816	8.88%
204088621	8.77%
308934168	6.82%
308809884	4.96%
308954480	4.24%
126257994	3.51%
204358234	3.02%
211245455	2.61%
230628900	1.72%
9541365	1.55%
99175835	1.46%
144229239	1.34%
9139673	1.28%
9706775	1.23%

Distance between 2 consecutive geolocations in 8 minutes intervals was approximated by Haversine formula and distance traveled during processing each field/municipality was estimated:

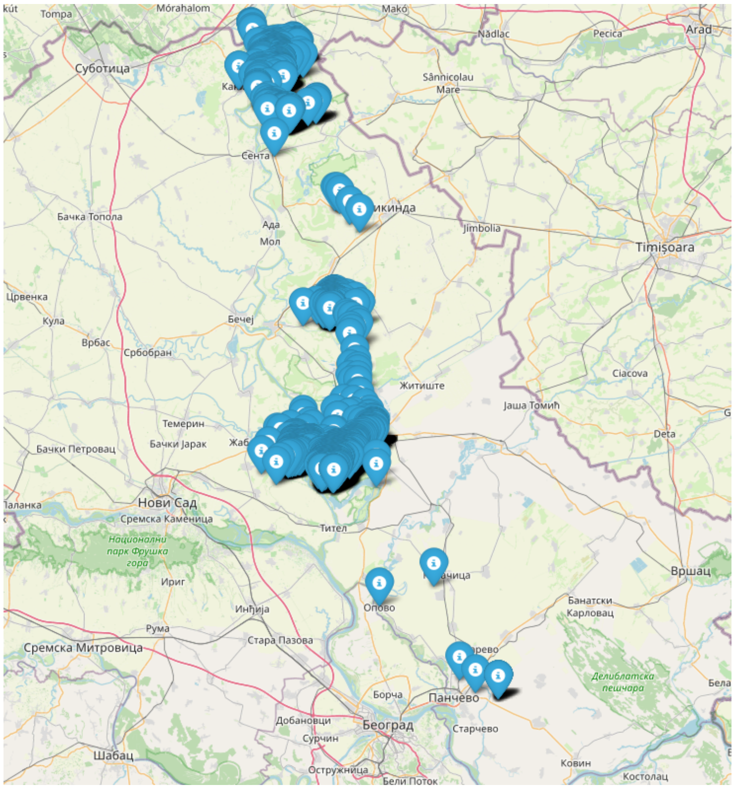


	municipality_distance_traveled
city	
Град Зрењанин	36.58 %
Општина Нови Кнежевац	35.05 %
Општина Чока	7.46 %
Град Кикинда	4.44 %
Општина Жабаљ	4.13 %
Општина Нови Бечеј	3.85 %
Општина Тител	2.47 %
Општина Ковачица	1.96 %
Општина Бечеј	0.81 %
Град Панчево	0.78 %
Општина Бачка Топола	0.41 %
Општина Сента	0.32 %
Општина Ада	0.3 %
Општина Темерин	0.3 %
Општина Кањижа	0.28 %
Град Сомбор	0.23 %
Општина Србобран	0.22 %
Град Нови Сад	0.21 %
Град Суботица	0.13 %
Општина Житиште	0.04 %
Општина Опово	0.04 %

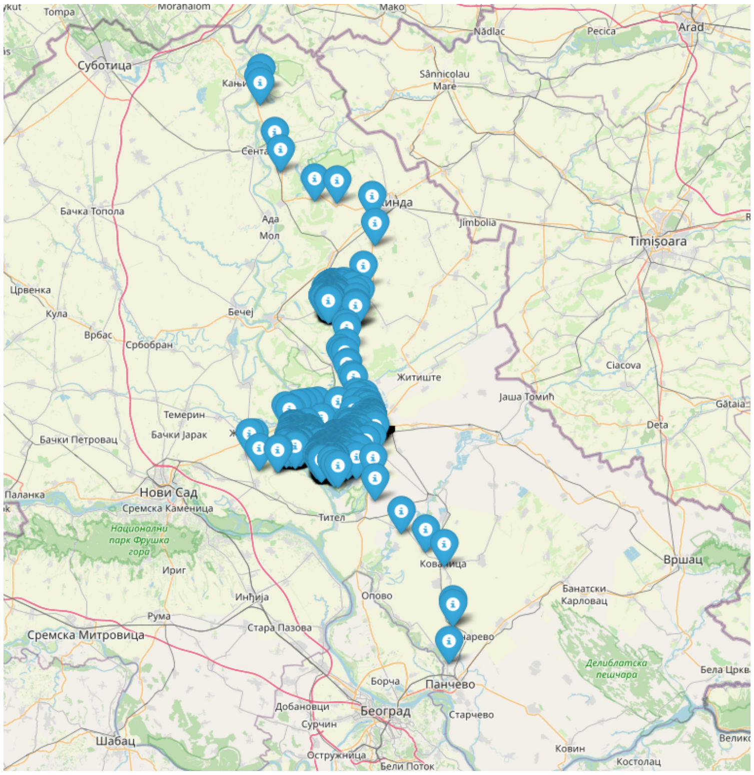


	field_distance_traveled
field_id	
204088621	10.82 %
309084405	8.83 %
308966765	6.94 %
308907816	6.27 %
211245455	3.97 %
308934168	2.73 %
144777455	2.43 %
308954480	2.37 %
308809884	2.18 %
147227440	1.85 %
9706775	1.37 %
204358234	1.3 %
210308907	1.28 %
211817889	1.26 %
147795929	1.2 %
144229239	1.13 %
9541365	1.11 %

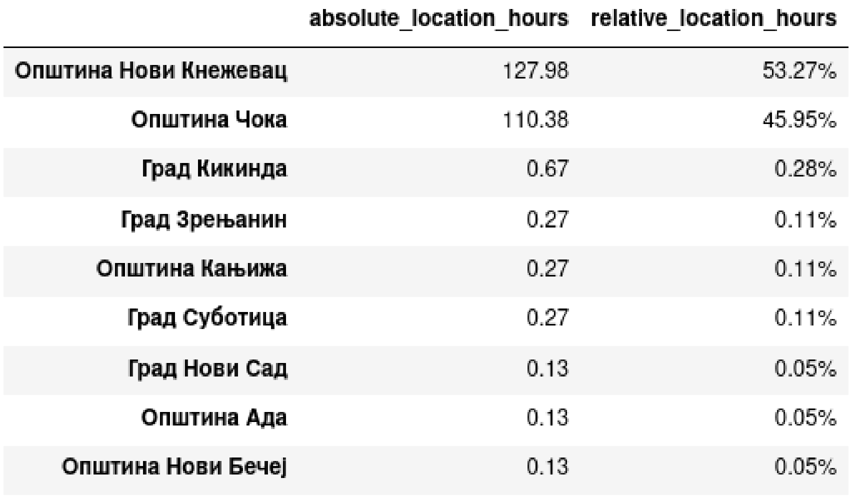
A2302895 sampled path:



A2302959 sampled path:

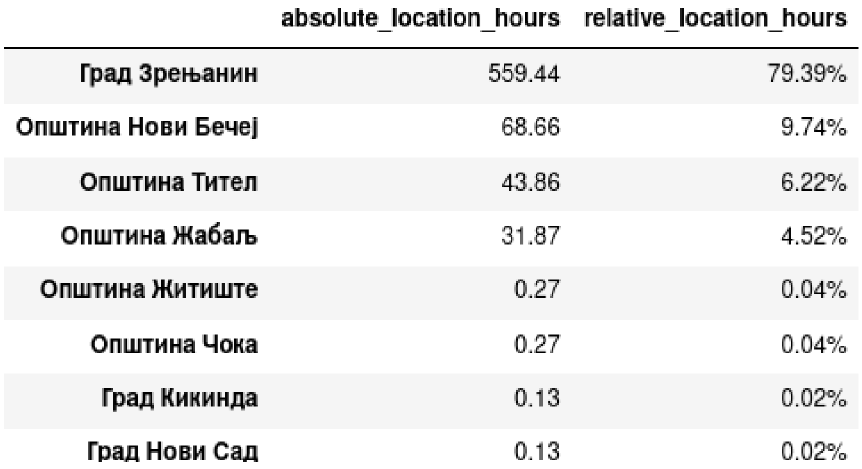


Machine A6002059, throughout its 300 days life span has approximated 240 total field working hours, with time allocation over 21 different field IDs, located mostly around Novi Knezevac and Coka:



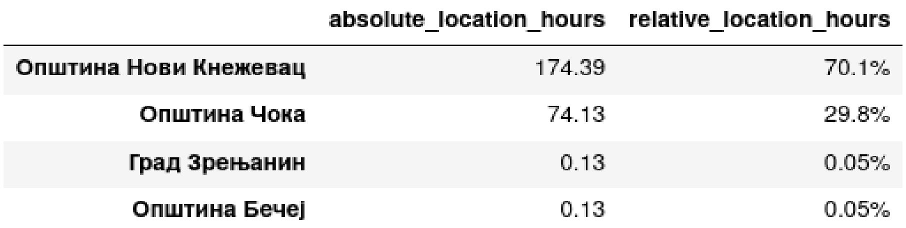
	absolute_location_hours	relative_location_hours
Општина Нови Кнежевац	127.98	53.27%
Општина Чока	110.38	45.95%
Град Кикинда	0.67	0.28%
Град Зрењанин	0.27	0.11%
Општина Кањижа	0.27	0.11%
Град Суботица	0.27	0.11%
Град Нови Сад	0.13	0.05%
Општина Ада	0.13	0.05%
Општина Нови Бечеј	0.13	0.05%

Machine A6002058, throughout its 304 days life span has approximated 705 total field working hours, with time allocation over 28 different field IDs, located mostly around Zrenjanin, Novi Becej, Titel and Zabalj:



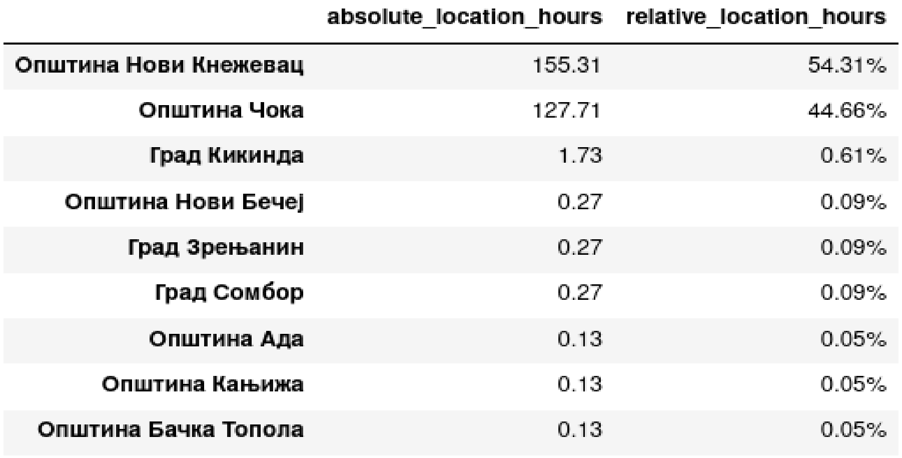
	absolute_location_hours	relative_location_hours
Град Зрењанин	559.44	79.39%
Општина Нови Бечеј	68.66	9.74%
Општина Тител	43.86	6.22%
Општина Жабаљ	31.87	4.52%
Општина Житиште	0.27	0.04%
Општина Чока	0.27	0.04%
Град Кикинда	0.13	0.02%
Град Нови Сад	0.13	0.02%

Machine A7702023, throughout its 339 days life span has approximated 249 total field working hours, with time allocation over 21 different field IDs, located mostly around Novi Knezevac and Coka:



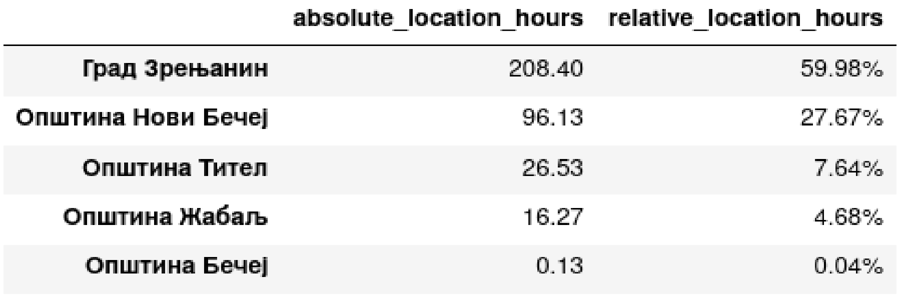
	absolute_location_hours	relative_location_hours
Општина Нови Кнежевац	174.39	70.1%
Општина Чока	74.13	29.8%
Град Зрењанин	0.13	0.05%
Општина Бечеј	0.13	0.05%

Machine A7702039, throughout its 305 days life span has approximated 286 total field working hours, with time allocation over 32 different field IDs, located mostly around Novi Knezevac (~54.5%) and Coka (~44.5%) and Novi Becej, Kikinda, Sombor, Ada, Kanjiza, Zrenjanin and Backa Topola with less than 1%:



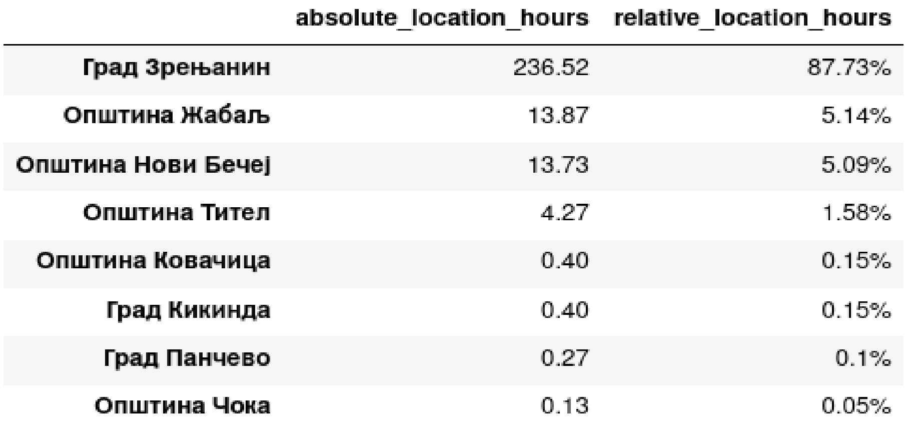
	absolute_location_hours	relative_location_hours
Општина Нови Кнежевац	155.31	54.31%
Општина Чока	127.71	44.66%
Град Кикинда	1.73	0.61%
Општина Нови Бечеј	0.27	0.09%
Град Зрењанин	0.27	0.09%
Град Сомбор	0.27	0.09%
Општина Ада	0.13	0.05%
Општина Кањижа	0.13	0.05%
Општина Бачка Топола	0.13	0.05%

Machine A7702039, throughout its 302 days life span has approximated 348 total field working hours, with time allocation over 18 different field IDs, located mostly around Zrenjanin, Novi Becej, Titel and Zabalj:



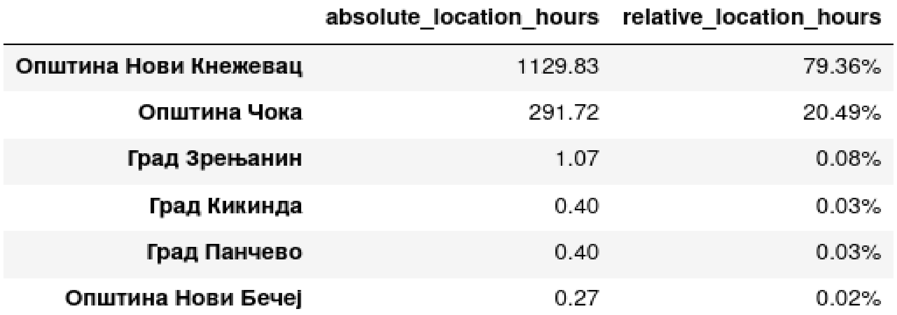
	absolute_location_hours	relative_location_hours
Град Зрењанин	208.40	59.98%
Општина Нови Бечеј	96.13	27.67%
Општина Тител	26.53	7.64%
Општина Жабаљ	16.27	4.68%
Општина Бечеј	0.13	0.04%

Machine A7702039, throughout its 224 days life span has approximated 270 total field working hours, with time allocation over 23 different field IDs, located mostly around Zrenjanin, Novi Becej, Zabalj and Titel:



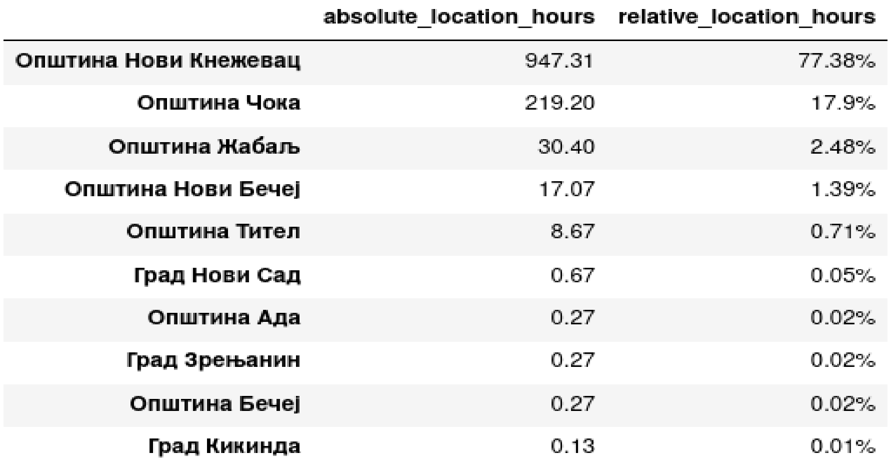
	absolute_location_hours	relative_location_hours
Град Зрењанин	236.52	87.73%
Општина Жабаљ	13.87	5.14%
Општина Нови Бечеј	13.73	5.09%
Општина Тител	4.27	1.58%
Општина Ковачица	0.40	0.15%
Град Кикинда	0.40	0.15%
Град Панчево	0.27	0.1%
Општина Чока	0.13	0.05%

Machine A2302888, throughout its 505 days life span has approximated 1424 total field working hours, with time allocation over 33 different field IDs, located mostly around Novi Knezevac and Coka:



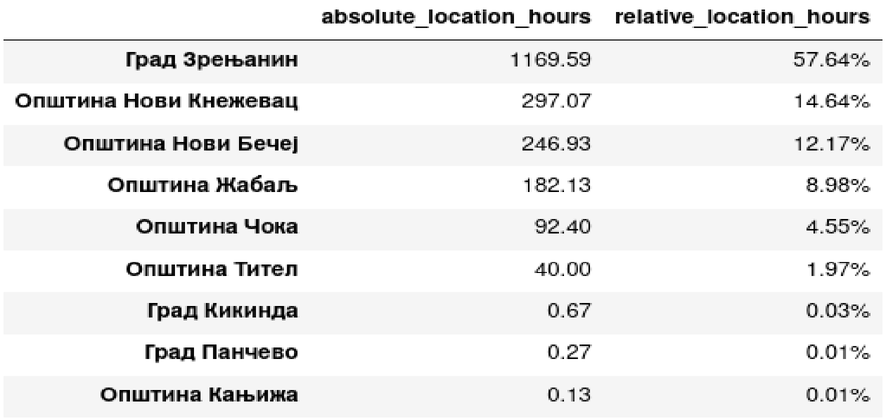
	absolute_location_hours	relative_location_hours
Општина Нови Кнежевац	1129.83	79.36%
Општина Чока	291.72	20.49%
Град Зрењанин	1.07	0.08%
Град Кикинда	0.40	0.03%
Град Панчево	0.40	0.03%
Општина Нови Бечеј	0.27	0.02%

Machine A2302888, throughout its 506 days life span has approximated 1224 total field working hours, with time allocation over 36 different field IDs, located mostly around Novi Knezevac, Coka, Zabalj and Novi Becej:



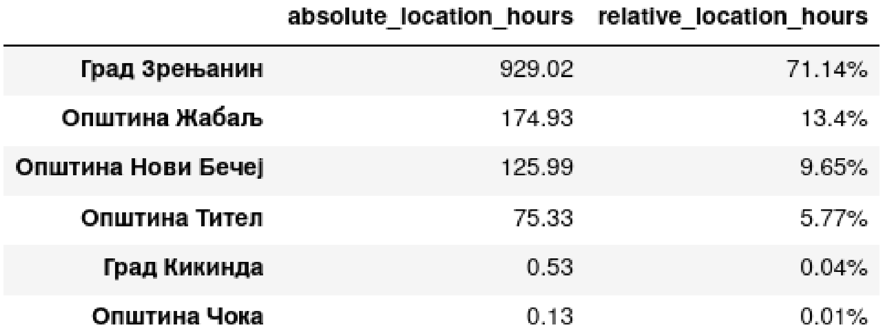
	absolute_location_hours	relative_location_hours
Општина Нови Кнежевац	947.31	77.38%
Општина Чока	219.20	17.9%
Општина Жабаљ	30.40	2.48%
Општина Нови Бечеј	17.07	1.39%
Општина Тител	8.67	0.71%
Град Нови Сад	0.67	0.05%
Општина Ада	0.27	0.02%
Град Зрењанин	0.27	0.02%
Општина Бечеј	0.27	0.02%
Град Кикинда	0.13	0.01%

Machine A2302895, throughout its 526 days life span has approximated 2029 total field working hours, with time allocation over 48 different field IDs, located mostly around Zrenjanin, Novi Knezevac, Novi Becej, Zabalj, Coka, and Titel:



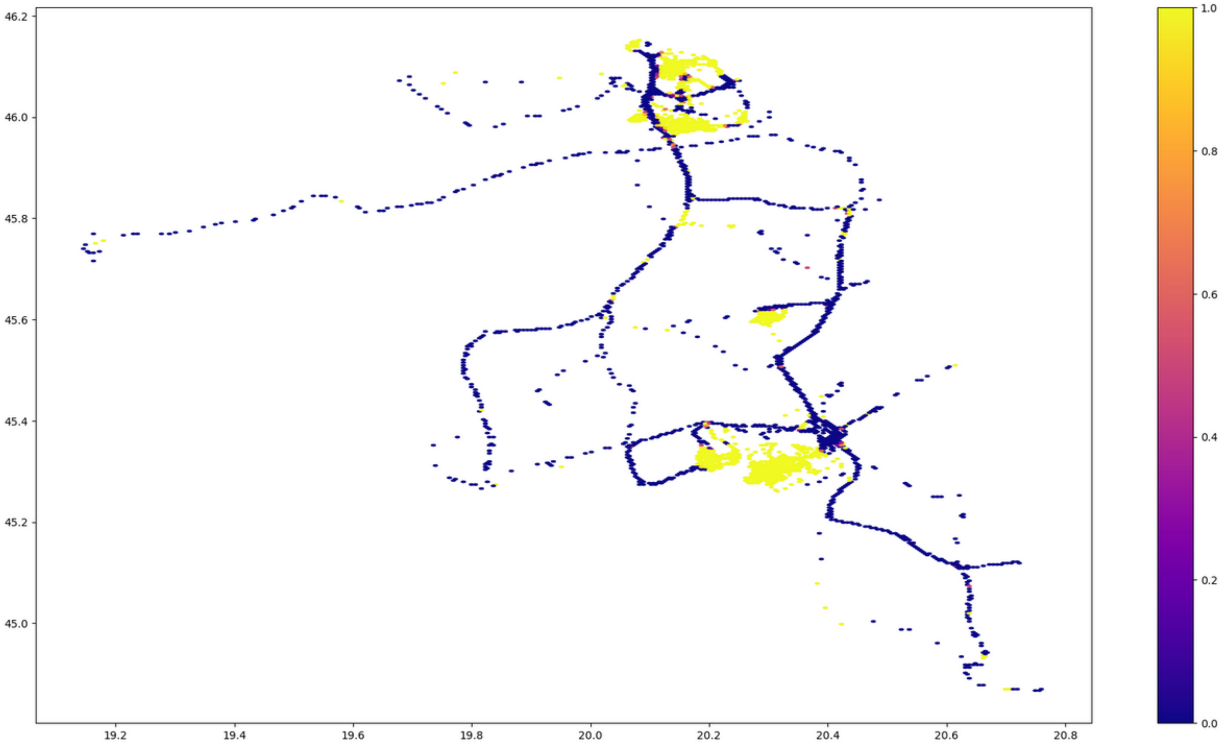
	absolute_location_hours	relative_location_hours
Град Зрењанин	1169.59	57.64%
Општина Нови Кнежевац	297.07	14.64%
Општина Нови Бечеј	246.93	12.17%
Општина Жабаљ	182.13	8.98%
Општина Чока	92.40	4.55%
Општина Тител	40.00	1.97%
Град Кикинда	0.67	0.03%
Град Панчево	0.27	0.01%
Општина Кањижа	0.13	0.01%

Machine A2302959, throughout its 404 days life span has approximated 1306 total field working hours, with time allocation over 34 different field IDs, located mostly around Zrenjanin, Zabalj, Novi Becej and Titel:



	absolute_location_hours	relative_location_hours
Град Зрењанин	929.02	71.14%
Општина Жабаљ	174.93	13.4%
Општина Нови Бечеј	125.99	9.65%
Општина Тител	75.33	5.77%
Град Кикинда	0.53	0.04%
Општина Чока	0.13	0.01%

Overall depiction of transportation/field operation ratio for all machines:



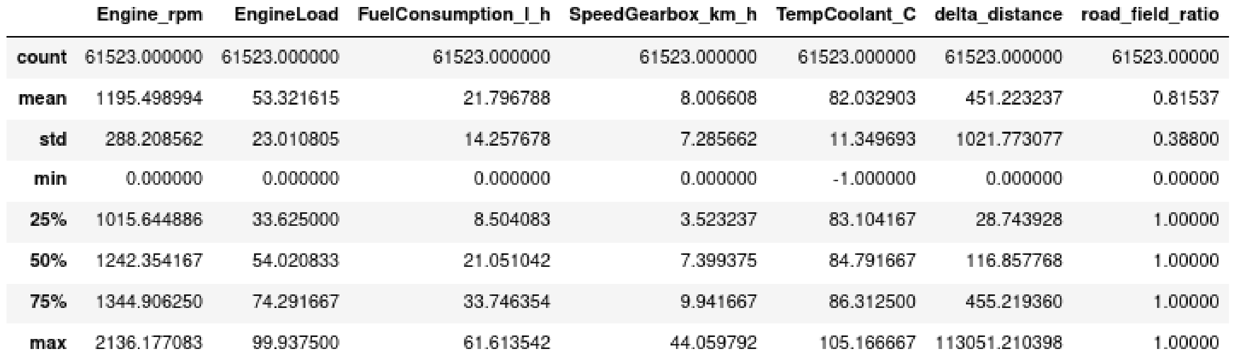
Yellow area represents field operations, whereas blue lines are estimated as transportation. As assumed earlier, there are several highly dense field clusters, and several smaller field areas slightly dislocated from these larger fields.

3. How many different types of field operations (based on telemetry data) did the mechanization execute?

3.1 Problem formulation:

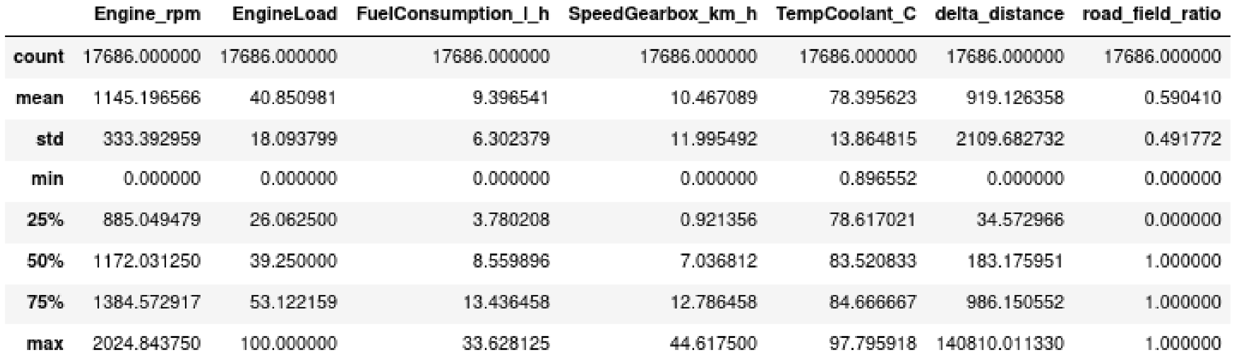
From the above findings on field hours analysis for each machine, in order to avoid trying to classify the operation work of different type of machines, data was preliminarily divided into 2 subsamples, where primary division criteria is the estimated number of hours, where each samples are composed out of 5 machines:

Sample 1 [A2302888, A2302900, A2302895, A2302959, A6002058] total operations summary statistics:



	Engine_rpm	EngineLoad	FuelConsumption_l_h	SpeedGearbox_km_h	TempCoolant_C	delta_distance	road_field_ratio
count	61523.000000	61523.000000	61523.000000	61523.000000	61523.000000	61523.000000	61523.00000
mean	1195.498994	53.321615	21.796788	8.006608	82.032903	451.223237	0.81537
std	288.208562	23.010805	14.257678	7.285662	11.349693	1021.773077	0.38800
min	0.000000	0.000000	0.000000	0.000000	-1.000000	0.000000	0.00000
25%	1015.644886	33.625000	8.504083	3.523237	83.104167	28.743928	1.00000
50%	1242.354167	54.020833	21.051042	7.399375	84.791667	116.857768	1.00000
75%	1344.906250	74.291667	33.746354	9.941667	86.312500	455.219360	1.00000
max	2136.177083	99.937500	61.613542	44.059792	105.166667	113051.210398	1.00000

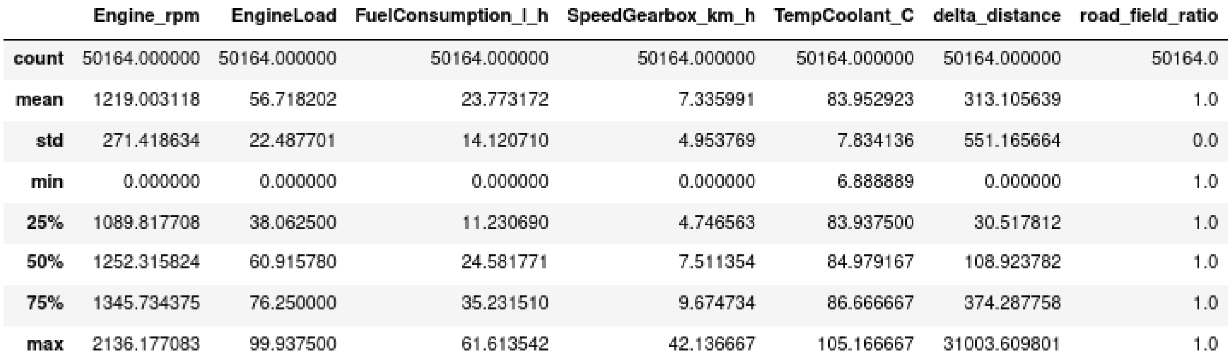
Sample 2 [A7702023, A6002058, A7702023, A7702039, A7702043, A7702047] total operations summary statistics:



	Engine_rpm	EngineLoad	FuelConsumption_l_h	SpeedGearbox_km_h	TempCoolant_C	delta_distance	road_field_ratio
count	17686.000000	17686.000000	17686.000000	17686.000000	17686.000000	17686.000000	17686.000000
mean	1145.196566	40.850981	9.396541	10.467089	78.395623	919.126358	0.590410
std	333.392959	18.093799	6.302379	11.995492	13.864815	2109.682732	0.491772
min	0.000000	0.000000	0.000000	0.000000	0.896552	0.000000	0.000000
25%	885.049479	26.062500	3.780208	0.921356	78.617021	34.572966	0.000000
50%	1172.031250	39.250000	8.559896	7.036812	83.520833	183.175951	1.000000
75%	1384.572917	53.122159	13.436458	12.786458	84.666667	986.150552	1.000000
max	2024.843750	100.000000	33.628125	44.617500	97.795918	140810.011330	1.000000

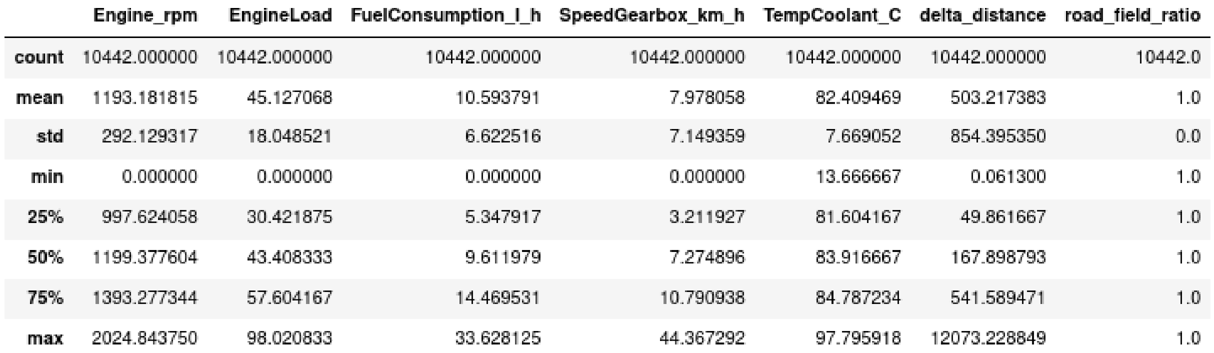
Besides substantial difference sample sizes expressed in operation hours, these 2 tables indicate that there is notable disparity between subsamples in terms of fuel consumption, approximated distance between 2 consecutive locations within an 8 minute interval and transportation to field ratio. Moreover, the higher levels of engine load and machine temperature are observed.

Sample 1 [A2302888, A2302900, A2302895, A2302959, A6002058] field operations summary statistics:



	Engine_rpm	EngineLoad	FuelConsumption_l_h	SpeedGearbox_km_h	TempCoolant_C	delta_distance	road_field_ratio
count	50164.000000	50164.000000	50164.000000	50164.000000	50164.000000	50164.000000	50164.0
mean	1219.003118	56.718202	23.773172	7.335991	83.952923	313.105639	1.0
std	271.418634	22.487701	14.120710	4.953769	7.834136	551.165664	0.0
min	0.000000	0.000000	0.000000	0.000000	6.888889	0.000000	1.0
25%	1089.817708	38.062500	11.230690	4.746563	83.937500	30.517812	1.0
50%	1252.315824	60.915780	24.581771	7.511354	84.979167	108.923782	1.0
75%	1345.734375	76.250000	35.231510	9.674734	86.666667	374.287758	1.0
max	2136.177083	99.937500	61.613542	42.136667	105.166667	31003.609801	1.0

Sample 2 [A7702023, A6002058, A7702023, A7702039, A7702043, A7702047] field operations summary statistics:



	Engine_rpm	EngineLoad	FuelConsumption_l_h	SpeedGearbox_km_h	TempCoolant_C	delta_distance	road_field_ratio
count	10442.000000	10442.000000	10442.000000	10442.000000	10442.000000	10442.000000	10442.0
mean	1193.181815	45.127068	10.593791	7.978058	82.409469	503.217383	1.0
std	292.129317	18.048521	6.622516	7.149359	7.669052	854.395350	0.0
min	0.000000	0.000000	0.000000	0.000000	13.666667	0.061300	1.0
25%	997.624058	30.421875	5.347917	3.211927	81.604167	49.861667	1.0
50%	1199.377604	43.408333	9.611979	7.274896	83.916667	167.898793	1.0
75%	1393.277344	57.604167	14.469531	10.790938	84.787234	541.589471	1.0
max	2024.843750	98.020833	33.628125	44.367292	97.795918	12073.228849	1.0

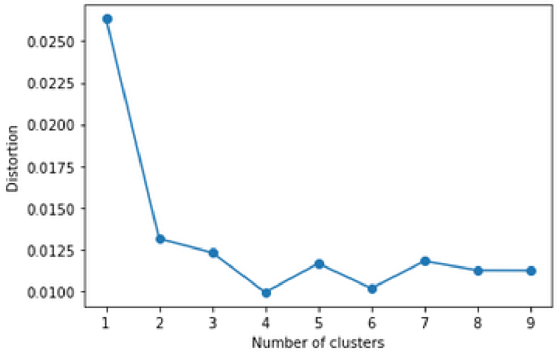
When analyzing solely field hours, the similar findings as for total operation hours could be observed.

3. 1 Method:

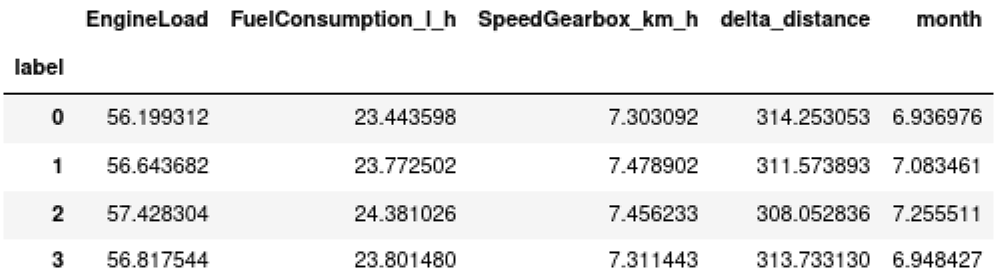
In order to estimate the number of field operations, the assumption regarding operational uniformity across the field is imposed, meaning that it is assumed that on each field similar operations are conducted (i.e. plowing, sowing, maintaining, harvesting…). Next, tslearn library and k-Shape multivariate time series clustering was performed to approximate and predict the cluster label for both subsamples, over the search space in range of 1-15 clusters. Clustering was performed over 5 selected model variables: load, fuel consumption, speed, delta distance and month of the year. Main advantages of this method is its scalability, generalization and easy implementation. However, some more informative geospatial features might be missing such as the delta change in terrain elevation or approximated field area, in order to increase explanatory capability of the model. In such cases, possible dimensionality reduction, for example PCA, could be implemented together with testing approaches from the spectral clustering algorithms realm.

3.3 Findings:

By applying the elbow technique to determine the number of clusters, for sample 1 with the machines of higher utilization levels, 4 clusters were estimated:



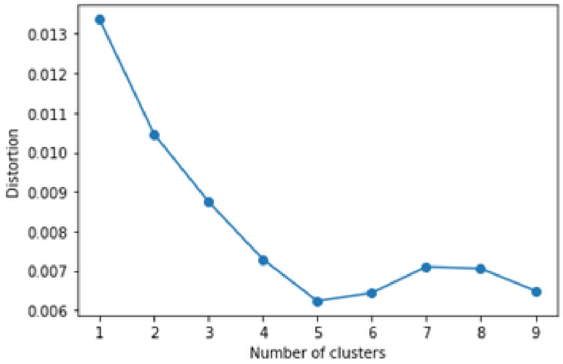
However, after inference was implemented and observations were labeled by cluster IDs, no significant variance in high level statistic across the clusters was observed:

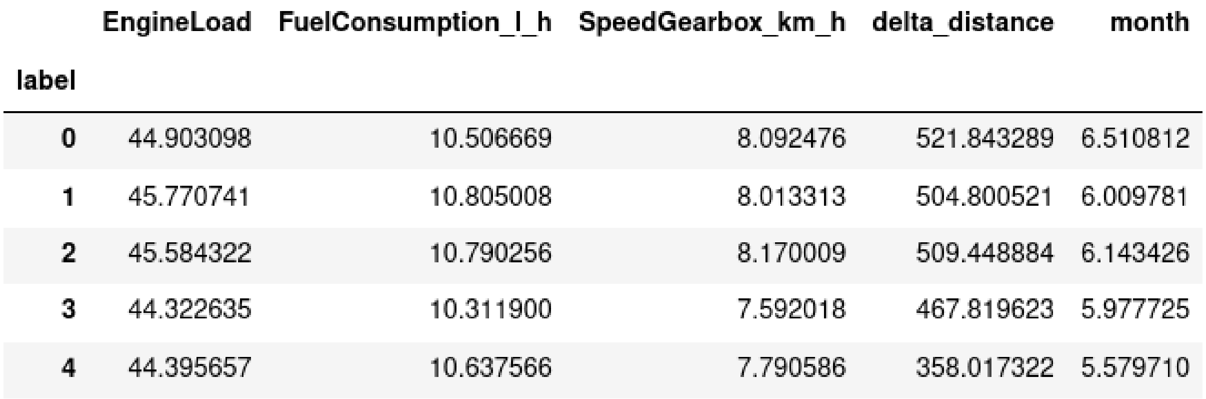


	EngineLoad	FuelConsumption_l_h	SpeedGearbox_km_h	delta_distance	month
label					
0	56.199312	23.443598	7.303092	314.253053	6.936976
1	56.643682	23.772502	7.478902	311.573893	7.083461
2	57.428304	24.381026	7.456233	308.052836	7.255511
3	56.817544	23.801480	7.311443	313.733130	6.948427

Cluster 3 accounted for the largest representation with 51% in the time series, followed by cluster 0 with 28% and clusters 1 and 2 with around 10%.

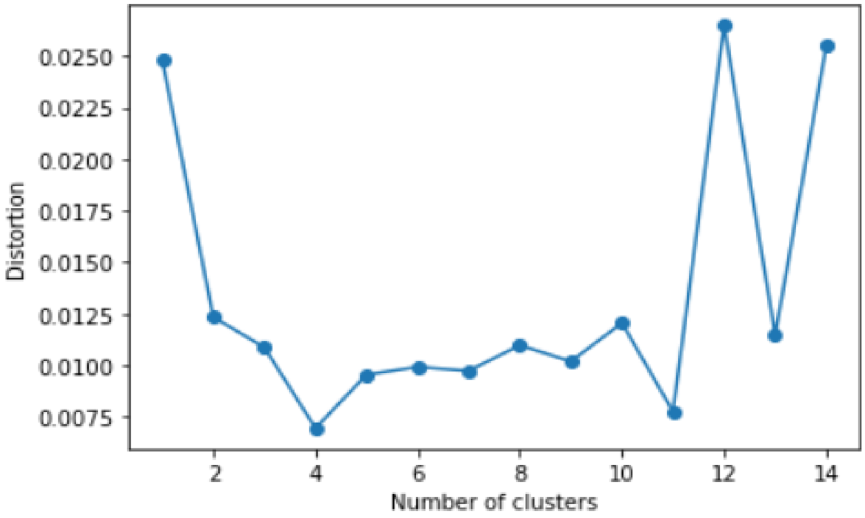
Regarding subsample 2, by applying the identical method 5 clusters were estimated:

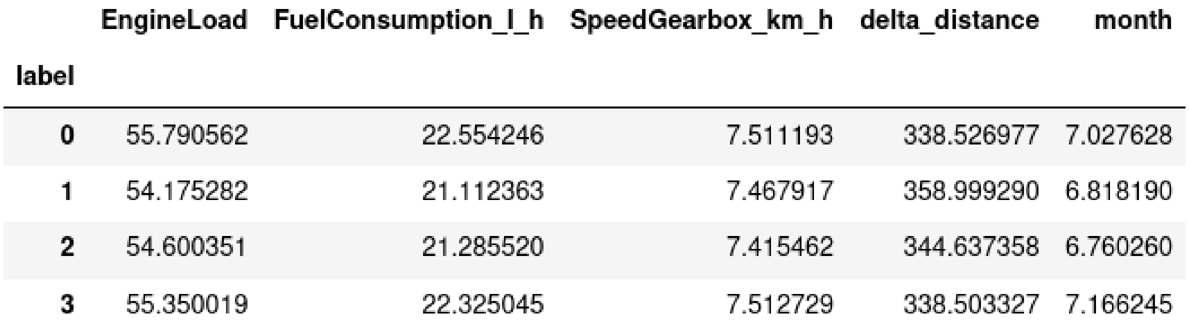




label	EngineLoad	FuelConsumption_l_h	SpeedGearbox_km_h	delta_distance	month
0	44.903098	10.506669	8.092476	521.843289	6.510812
1	45.770741	10.805008	8.013313	504.800521	6.009781
2	45.584322	10.790256	8.170009	509.448884	6.143426
3	44.322635	10.311900	7.592018	467.819623	5.977725
4	44.395657	10.637566	7.790586	358.017322	5.579710

When analyzing summary statistics and distributions for the second subsample, although 5 clusters were identified, one cluster had negligible small representation with less than 0.5%. In order to have reliable inference, another round of clustering for individual machines from subsample 2 was performed. When analyzing individual plots, it was observed that for most of the machines, the distortion curve was decreasing steadily until 4 clusters, then it was oscillating around the same level of distortion with small marginal improvement with increasing number of clusters. The next convergence was between 8-10 clusters. Therefore, the final round clustering was performed on overall sample and again, 4 clusters were identified:



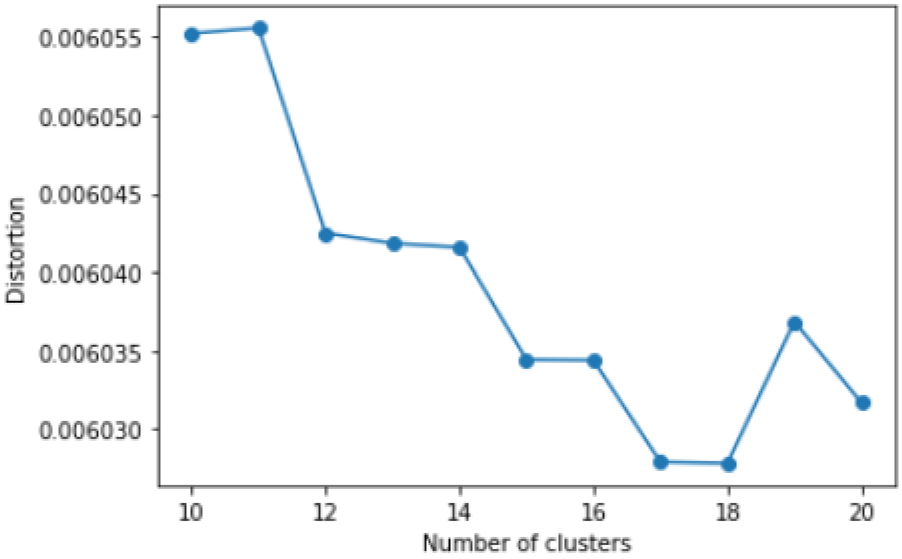


label	EngineLoad	FuelConsumption_l_h	SpeedGearbox_km_h	delta_distance	month
0	55.790562	22.554246	7.511193	338.526977	7.027628
1	54.175282	21.112363	7.467917	358.999290	6.818190
2	54.600351	21.285520	7.415462	344.637358	6.760260
3	55.350019	22.325045	7.512729	338.503327	7.166245

Regarding individual clusters’ representation, cluster 2 was most dominant with 57% of time allocated to this activity, followed by cluster 3 with 20%, cluster 1 with 17% and cluster 0 with 6%. This structure could confirm the assumption regarding 4 basic field operations such as plowing, sowing, harvesting and field maintaining. However, the difference in some variables’ values of starting 2 subsamples could have been explained by machines of different types (i.e. sizes and powers - explains the difference in consumption and load) working on fields of different sizes (machines from subsample 2 might have been working primarily on smaller fields since delta distance is larger). Also, since for subsample 2, the optimal number of clusters could be higher (8-10), these machines might have been employed in some more specialized operations (spraying, grass removal…)

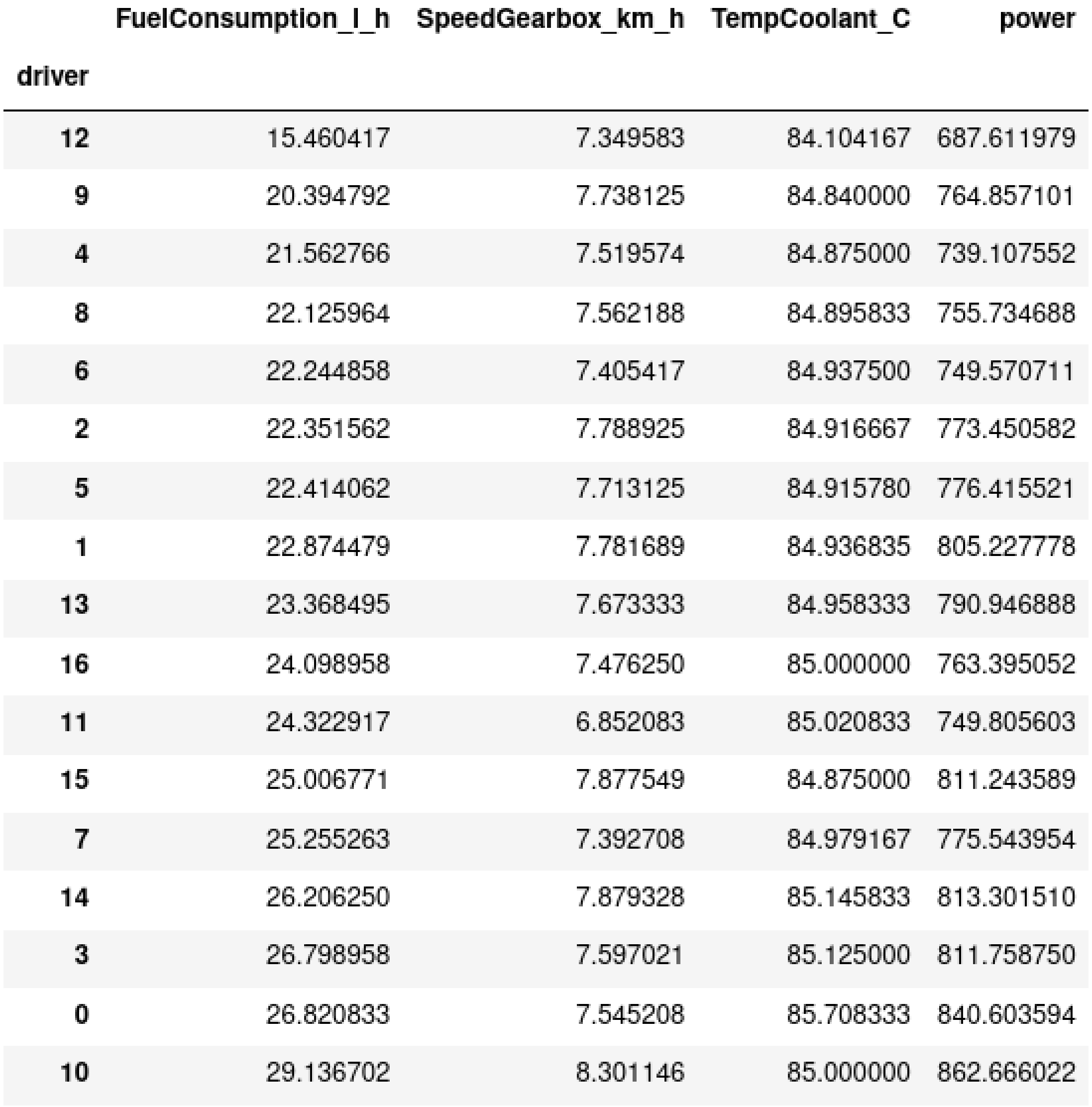
4. Can you estimate how many drivers are operating the fleet of machines?

Similarly to the previous problem, multivariate model based clustering was performed. In order to represent each machine’s power utilization, the engine's revolutions per minute were multiplied by the engine’s loads. Additionally, the average speed and temperature were included. Model assumes that each driver can drive any machine. By this set up, the model is aiming to identify how much of each machine’s relative power and fuel is consumed to execute the same task - operation with the cluster label 2 which accounts for more than 57% of observations. Moreover, to ensure a comparable time period and “stable” fleet size, the observed period was sampled down to the year 2019 and months from July up to and including November. By such approach 17 or 18 drivers were estimated:



5. Does field consumption differ between different drivers?

By analyzing median values of model variables, following values are observed:



	FuelConsumption_l_h	SpeedGearbox_km_h	TempCoolant_C	power
driver				
12	15.460417	7.349583	84.104167	687.611979
9	20.394792	7.738125	84.840000	764.857101
4	21.562766	7.519574	84.875000	739.107552
8	22.125964	7.562188	84.895833	755.734688
6	22.244858	7.405417	84.937500	749.570711
2	22.351562	7.788925	84.916667	773.450582
5	22.414062	7.713125	84.915780	776.415521
1	22.874479	7.781689	84.936835	805.227778
13	23.368495	7.673333	84.958333	790.946888
16	24.098958	7.476250	85.000000	763.395052
11	24.322917	6.852083	85.020833	749.805603
15	25.006771	7.877549	84.875000	811.243589
7	25.255263	7.392708	84.979167	775.543954
14	26.206250	7.879328	85.145833	813.301510
3	26.798958	7.597021	85.125000	811.758750
0	26.820833	7.545208	85.708333	840.603594
10	29.136702	8.301146	85.000000	862.666022

This approach indicates differences in consumption between the drivers in terms of consumption, power and speed but not in terms of the temperature. However, in order to have precise insight into each driver’s capabilities, the machine types or IDs should have been incorporated into the model, either as individual model variable or weighting factor for all variables.

6. Can you estimate what would be the optimal way of driving for some tractor on some plot (is there a general way to determine, for any combination of driver/tractor/plot)?

Theoretically, different optimization strategies could be developed for a given task, i.e. by combining heuristics based on analytics' findings together with more complex routing optimization algorithms. Since field operations, from time perspective, can overlap for different agricultural crops, at the first step, it should be decided if optimization is performed for each operation separately or all operations are treated equally. Moreover, besides costs, there should be some time frame constraints imposed for each field or operation or field/operation combination.

Moreover, If there is no established field prioritization system, then for each field, the number of working hours could be used for cost function approximation and this information could be used to estimate the number of drivers needed to process each field. For each field, drivers ranking could be calculated by sorting the driver's average fuel consumption for a given field id. Given all field IDs and corresponding drivers, for each driver a start and end nodes together with distance matrix could be calculated and some of the vehicle route optimization algorithms, i.e. simple min cost flow program could be implemented for each driver individually or some more complex form of multiple traveling salesman problem for all drivers in bulk.

Appendix

Figure 1: Estimated delta distance over 8 minutes time windows. Purple hexagons indicate field work whereas yellow ones indicate transport.

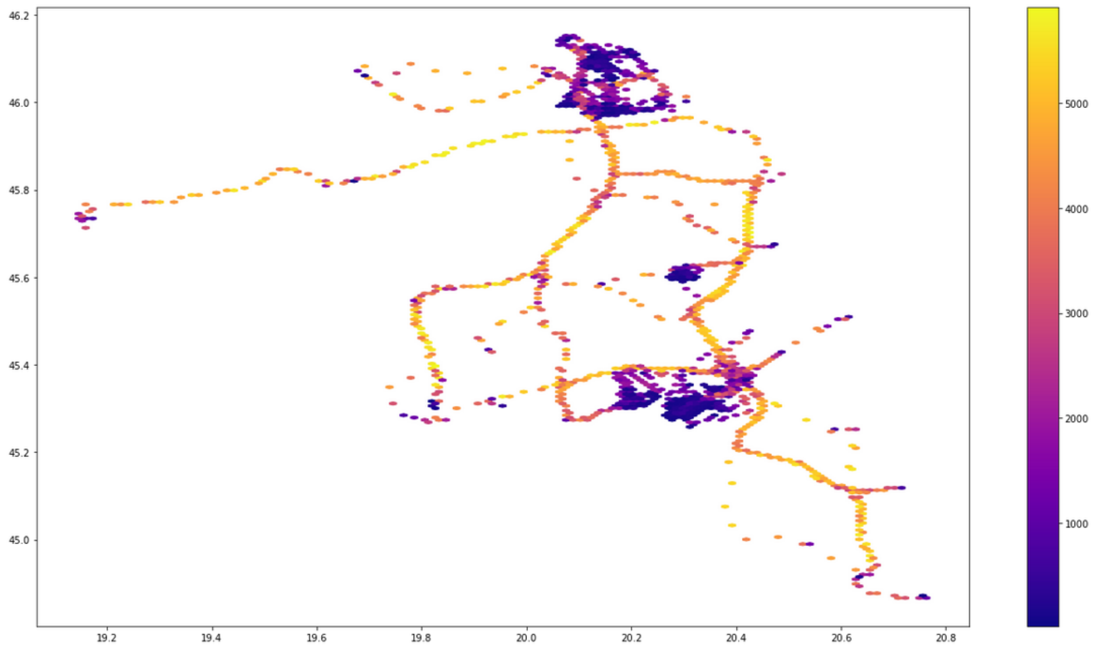


Figure 2: Estimated average fuel consumption over 8 minutes time windows. Purple hexagons indicate fields with lower consumptions whereas the highest consumption is observed close to the largest field clusters.

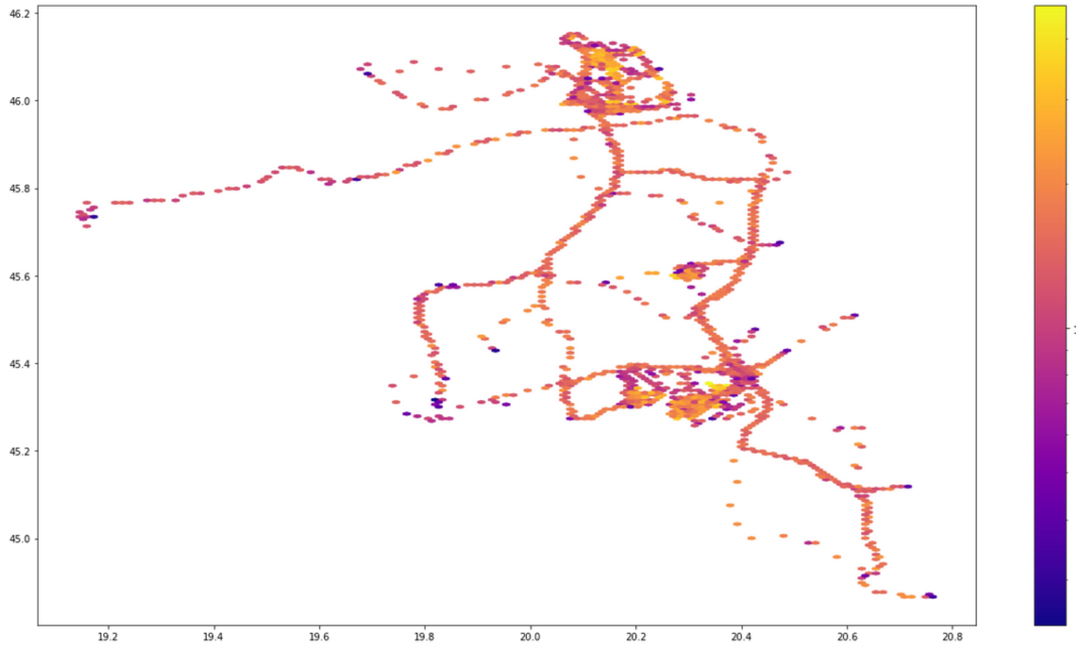
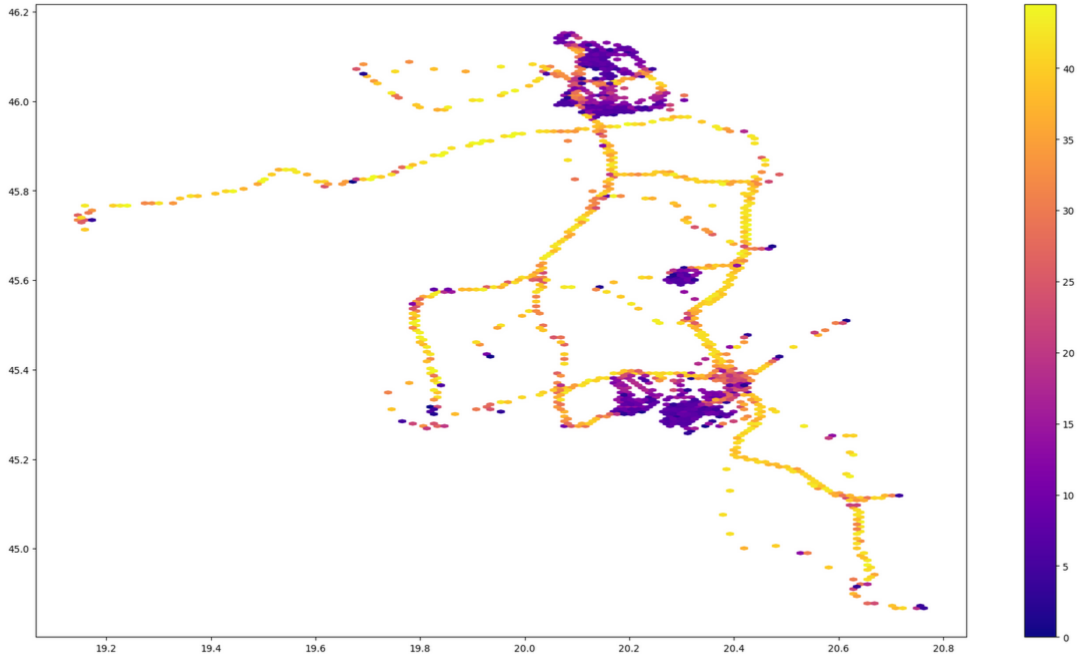


Figure 3: Estimated average speed over 8 minutes time windows. Purple hexagons indicate lower speed levels whereas the highest speed is observed close to road areas.


46.2
46.0
45.8
45.6
45.4
45.2
45.0
19.2
19.4
19.6
19.8
20.0
20.2
20.4
20.6
20.8
40
35
30
25
20
15
10
5
0