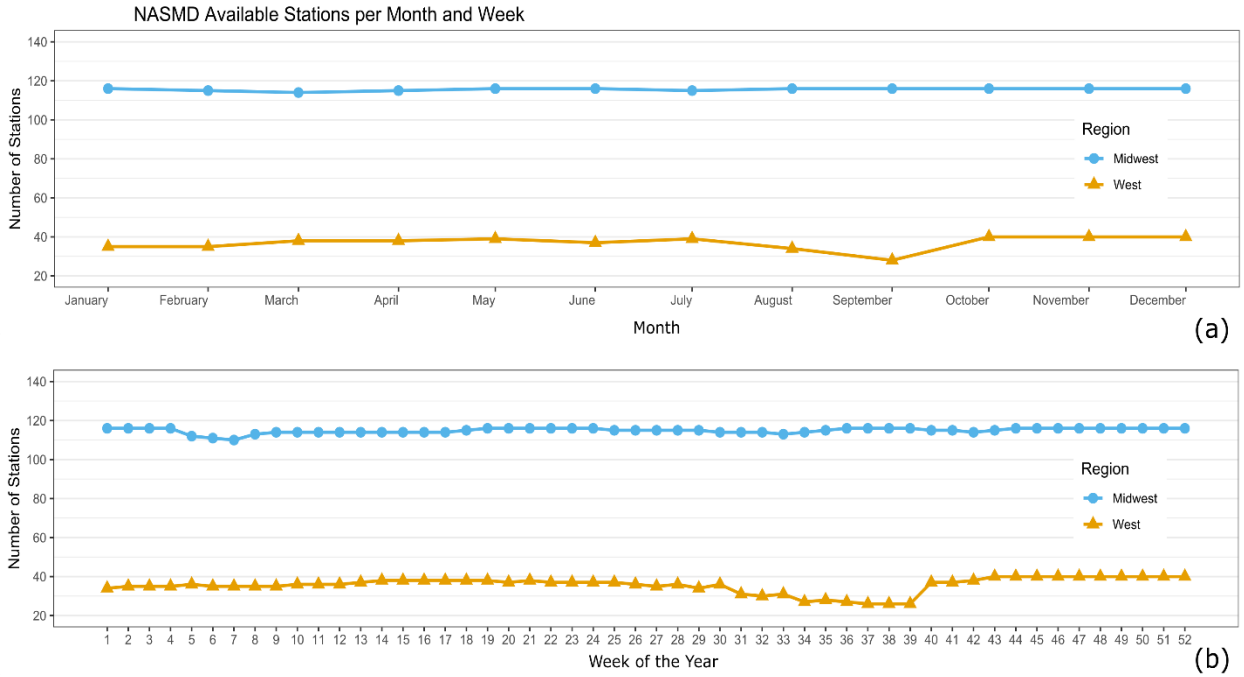
Downscaling Satellite Soil Moisture using a Modular Spatial Inference Framework

Ricardo M. Llamas, Leobardo Valera, Paula Olaya, Michela Taufer and Rodrigo Vargas

Supplementary Material S2

Number of North American Soil Moisture Database available stations in 2010 over the two regions of interest

The maximum number of NASMD stations available per month in the West region ranged from 28 to 39, while they ranged from 114 to 116 in the Midwest region. Similar numbers of stations were available regarding weekly periods, ranging from 26 to 39 in the West region, and 110 to 116 in the Midwest region.



Output parameters in model generation process for the four proposed machine learning methods

1. KKNN model generation outputs for monthly layers in 2010 over the Midwest region of interest

The **Layer** column shows the month of the year. The **No.Points** column shows the number of points used as training data in the entire sampling space in every month, this number corresponds with the 70% of the available sample points that were defined for training the model. The **kmax** column shows the maximum threshold of k-neighbors allowed to generate each model. The **Ktunned** column shows the optimal number of k-neighbors in each model derived from the highest correlation and lower rmse obtained from a 10-fold

cross validation process. The **Kernel** column shows the optimal kernel function obtained from a 10-fold cross validation process. The **rmse** and **Cor** columns show the best results of the root minimum square error and the correlation respectively, derived from a 10-fold cross validation process.

	Layer	No.Points	kmax	Ktunned	Kernel	rmse	Cor
1	2010_01	312	50	16	optimal	0.033384	0.842153
2	2010_02	314	50	22	triangular	0.034365	0.754509
3	2010_03	314	50	21	gaussian	0.03449	0.789885
4	2010_04	314	50	14	triangular	0.030992	0.818085
5	2010_05	314	50	17	triangular	0.032319	0.763043
6	2010_06	314	50	10	gaussian	0.032614	0.744214
7	2010_07	314	50	22	triangular	0.034059	0.619679
8	2010_08	314	50	14	gaussian	0.030717	0.589444
9	2010_09	314	50	16	optimal	0.031479	0.766476
10	2010_10	314	50	10	rectangular	0.029734	0.65627
11	2010_11	314	50	13	gaussian	0.033273	0.70344
12	2010_12	313	50	11	gaussian	0.03322	0.751877

2. KKNN model generation outputs for monthly layers in 2010 over the West region of interest.

The **Layer** column shows the month of the year. The **No.Points** column shows the number of points used as training data in the entire sampling space in every month, this number corresponds with the 70% of the available sample points that were defined for training the model. The **kmax** column shows the maximum threshold of k-neighbors allowed to generate each model. The **Ktunned** column shows the optimal number of k-neighbors in each model derived from the highest correlation and lower RMSE obtained from a 10-fold cross validation process. The **Kernel** column shows the optimal kernel function obtained from a 10-fold cross validation process. The **rmse** and **Cor** columns show the best results of the root minimum square error and the correlation respectively, derived from a 10-fold cross validation process.

	Layer	No.Points	kmax	Ktunned	Kernel	rmse	Cor
1	2010_01	147	50	9	epanechnikov	0.042735184	0.463605999
2	2010_02	170	50	12	optimal	0.042436006	0.470224521
3	2010_03	165	50	17	triangular	0.037342116	0.595236665
4	2010_04	254	50	9	gaussian	0.041393036	0.568183368
5	2010_05	279	50	17	gaussian	0.033858341	0.5571332
6	2010_06	306	50	18	epanechnikov	0.02704039	0.529970108
7	2010_07	301	50	12	gaussian	0.024741897	0.636102857
8	2010_08	299	50	25	gaussian	0.026257011	0.524157052
9	2010_09	275	50	19	optimal	0.024934333	0.491592833
10	2010_10	311	50	8	triangular	0.027085603	0.675251402
11	2010_11	307	50	15	epanechnikov	0.032414744	0.617932416
12	2010_12	249	50	9	gaussian	0.045342981	0.419021076

3. KKNN model generation outputs for weekly layers in 2010 over the Midwest region of interest.

The **Layer** column shows the week of the year based on the first Julian day of each given week. The **No.Points** column shows the number of points used as training data in the entire sampling space in every week, this number corresponds with the 70% of the available sample points that were defined for training the model. The **kmax** column shows the maximum threshold of k-neighbors allowed to generate each model. The **Ktunned** column shows the optimal number of k-neighbors in each model derived from the highest correlation and lower RMSE obtained from a 10-fold cross validation process. The **Kernel** column shows the optimal kernel function obtained from a 10-fold cross validation process. The **rmse** and **Cor** columns show the best results of the root minimum square error and the correlation respectively, derived from a 10-fold cross validation process.

	Layer	No.Points	kmax	Ktunned	Kernel	rmse	Cor
1	2010_001	259	50	15	gaussian	0.039293	0.756716
2	2010_008	269	50	9	gaussian	0.036164	0.777565
3	2010_015	309	50	14	triangular	0.034217	0.85282
4	2010_022	308	50	18	optimal	0.036187	0.820978
5	2010_029	289	50	14	optimal	0.046401	0.689352
6	2010_036	309	50	18	gaussian	0.033453	0.712355
7	2010_043	311	50	12	triangular	0.035511	0.74211
8	2010_050	312	50	14	triangular	0.032108	0.807246
9	2010_057	314	50	14	gaussian	0.032259	0.789468
10	2010_064	313	50	7	triangular	0.034287	0.793075
11	2010_071	314	50	13	triangular	0.033915	0.781181
12	2010_078	314	50	12	gaussian	0.034845	0.778537
13	2010_085	313	50	13	gaussian	0.033551	0.853531
14	2010_092	313	50	13	triangular	0.035666	0.894222
15	2010_099	313	50	13	gaussian	0.032455	0.881955
16	2010_106	314	50	6	gaussian	0.033087	0.597731
17	2010_113	313	50	15	triangular	0.032566	0.688552
18	2010_120	313	50	12	triangular	0.033301	0.825448
19	2010_127	314	50	9	triangular	0.033495	0.807414
20	2010_134	314	50	11	triangular	0.033487	0.688997
21	2010_141	313	50	12	triangular	0.032202	0.714525
22	2010_148	313	50	8	triangular	0.032294	0.736611
23	2010_155	314	50	15	gaussian	0.031964	0.81092
24	2010_162	314	50	15	optimal	0.034988	0.642114
25	2010_169	314	50	15	gaussian	0.032313	0.75854
26	2010_176	313	50	13	triangular	0.032719	0.682974
27	2010_183	314	50	10	triangular	0.032866	0.741059
28	2010_190	313	50	12	triangular	0.034006	0.621209
29	2010_197	314	50	8	triangular	0.033602	0.710816
30	2010_204	313	50	11	gaussian	0.032935	0.596883

31	2010_211	314	50	9	triangular	0.030585	0.618225
32	2010_218	313	50	11	gaussian	0.031838	0.5596
33	2010_225	314	50	19	triangular	0.035391	0.513209
34	2010_232	314	50	7	gaussian	0.034095	0.617448
35	2010_239	313	50	10	gaussian	0.032726	0.663503
36	2010_246	313	50	6	rank	0.031581	0.800421
37	2010_253	314	50	12	triangular	0.032768	0.814257
38	2010_260	313	50	9	optimal	0.033683	0.74188
39	2010_267	313	50	9	triangular	0.032853	0.693094
40	2010_274	314	50	12	optimal	0.031263	0.625529
41	2010_281	313	50	17	triangular	0.030474	0.751651
42	2010_288	314	50	19	gaussian	0.030446	0.790818
43	2010_295	313	50	10	rectangular	0.032593	0.489495
44	2010_302	313	50	14	gaussian	0.030825	0.730025
45	2010_309	314	50	9	gaussian	0.031649	0.733617
46	2010_316	313	50	12	triangular	0.033499	0.640483
47	2010_323	313	50	23	gaussian	0.034202	0.725369
48	2010_330	310	50	13	gaussian	0.032871	0.809627
49	2010_337	308	50	16	triangular	0.034715	0.778554
50	2010_344	298	50	29	gaussian	0.034828	0.742494
51	2010_351	269	50	17	gaussian	0.038912	0.619098
52	2010_358	235	50	18	triangular	0.03904	0.736184

4. KKNN model generation outputs for weekly layers in 2010 over the West region of interest.

The **Layer** column shows the week of the year based on the first Julian day of each given week. The **No.Points** column shows the number of points used as training data in the entire sampling space in every week, this number corresponds with the 70% of the available sample points that were defined for training the model. The **kmax** column shows the maximum threshold of k-neighbors allowed to generate each model. The **Ktunned** column shows the optimal number of k-neighbors in each model derived from the highest correlation and lower RMSE obtained from a 10-fold cross validation process. The **Kernel** column shows the optimal kernel function obtained from a 10-fold cross validation process. The **rmse** and **Cor** columns show the best results of the root minimum square error and the correlation respectively, derived from a 10-fold cross validation process.

	Layer	**No.Points**	**kmax**	**Ktunned**	**Kernel**	**rmse**	**Cor**
1	2010_001	82	50	5	epanechnikov	0.041601	0.71011
2	2010_008	92	50	5	gaussian	0.041825	0.641207
3	2010_015	134	50	8	rectangular	0.043103	0.477662
4	2010_022	110	50	7	gaussian	0.037851	0.712641
5	2010_029	79	50	9	triangular	0.044873	0.499388
6	2010_036	163	50	3	rectangular	0.042025	0.525727

7	2010_043	112	50	7	rectangular	0.041076	0.513411
8	2010_050	125	50	6	gaussian	0.037012	0.614376
9	2010_057	118	50	6	triangular	0.037202	0.785318
10	2010_064	144	50	9	triangular	0.033926	0.663868
11	2010_071	116	50	4	rank	0.038606	0.778592
12	2010_078	98	50	29	gaussian	0.044117	0.509771
13	2010_085	112	50	12	triangular	0.038498	0.544737
14	2010_092	136	50	11	triangular	0.041993	0.646883
15	2010_099	139	50	11	triangular	0.038531	0.655275
16	2010_106	213	50	14	optimal	0.042966	0.635265
17	2010_113	245	50	6	triangular	0.038932	0.660737
18	2010_120	185	50	11	triangular	0.039207	0.527385
19	2010_127	212	50	12	triangular	0.035879	0.659577
20	2010_134	271	50	17	gaussian	0.031118	0.578591
21	2010_141	232	50	6	rank	0.039611	0.621008
22	2010_148	263	50	8	rectangular	0.032489	0.677565
23	2010_155	282	50	16	rectangular	0.035451	0.523236
24	2010_162	277	50	18	gaussian	0.032029	0.511819
25	2010_169	183	50	25	gaussian	0.031265	0.316594
26	2010_176	224	50	45	gaussian	0.030587	0.405565
27	2010_183	213	50	17	gaussian	0.027988	0.436465
28	2010_190	228	50	18	rectangular	0.028274	0.2438
29	2010_197	273	50	32	gaussian	0.028025	0.329735
30	2010_204	266	50	9	gaussian	0.030123	0.683509
31	2010_211	250	50	8	gaussian	0.027597	0.591473
32	2010_218	226	50	9	rectangular	0.027893	0.619235
33	2010_225	230	50	24	gaussian	0.029999	0.526258
34	2010_232	235	50	8	gaussian	0.025714	0.600546
35	2010_239	283	50	41	gaussian	0.026607	0.427075
36	2010_246	244	50	18	triangular	0.029593	0.592949
37	2010_253	232	50	12	gaussian	0.026128	0.452776
38	2010_260	233	50	13	gaussian	0.027583	0.338615
39	2010_267	225	50	16	optimal	0.028581	0.442803
40	2010_274	310	50	11	epanechnikov	0.030875	0.627412
41	2010_281	308	50	8	gaussian	0.030789	0.734699
42	2010_288	310	50	6	triangular	0.032596	0.63296
43	2010_295	310	50	11	gaussian	0.032182	0.538848
44	2010_302	295	50	12	rectangular	0.033158	0.617948
45	2010_309	296	50	8	gaussian	0.034256	0.544957
46	2010_316	243	50	15	triangular	0.034164	0.59176
47	2010_323	203	50	10	triangular	0.039842	0.735292

48	2010_330	161	50	11	triangular	0.04244	0.628606
49	2010_337	179	50	49	gaussian	0.055432	0.414545
50	2010_344	171	50	8	triangular	0.032141	0.776669
51	2010_351	207	50	9	gaussian	0.037175	0.392222
52	2010_358	168	50	6	rectangular	0.042378	0.543952

5. **Random Forest model generation outputs for monthly layers in 2010 over the Midwest region of interest.**

The **Layer** column shows the month of the year. The **No.Points** column shows the number of points used as training data in the entire sampling space in every month, this number corresponds with the 70% of the available sample points that were defined for training the model. The **Best_Mtry** column shows the number of covariates defined in the model in relation to the number of predictors used as input (6 predictors for this study: latitude, longitude, elevation, aspect, slope and topographic wetness index). The **Trees** column shows the maximum number of trees allows in the model generation. The **rmse** and **Cor** columns show the best results of the root minimum square error and the correlation respectively, derived from a cross validation process.

	Layer	No.Points	Best_Mtry	Trees	rmse	Cor
1	2010_01	312	4	500	0.029525	0.874631
2	2010_02	314	2	500	0.032544	0.790907
3	2010_03	314	2	500	0.032361	0.825387
4	2010_04	314	4	500	0.028056	0.856964
5	2010_05	314	6	500	0.030829	0.792975
6	2010_06	314	6	500	0.030191	0.785509
7	2010_07	314	4	500	0.032083	0.679189
8	2010_08	314	4	500	0.028708	0.661906
9	2010_09	314	4	500	0.029513	0.798924
10	2010_10	314	2	500	0.027911	0.716726
11	2010_11	314	4	500	0.029857	0.772466
12	2010_12	313	4	500	0.030687	0.797773

6. **Random Forest model generation outputs for monthly layers in 2010 over the West region of interest.**

The **Layer** column shows the month of the year. The **No.Points** column shows the number of points used as training data in the entire sampling space in every month, this number corresponds with the 70% of the available sample points that were defined for training the model. The **Best_Mtry** column shows the number of covariates defined in the model in relation to the number of predictors used as input (6 predictors for this study: latitude, longitude, elevation, aspect, slope and topographic wetness index). The **Trees** column shows the maximum number of trees allows in the model generation. The **rmse** and **Cor** columns show the best results of the root minimum square error and the correlation respectively, derived from a cross validation process.

	Layer	No.Points	Best_Mtry	Trees	rmse	Cor
1	2010_01	147	2	500	0.040574	0.578781
2	2010_02	170	4	500	0.042011	0.511476
3	2010_03	165	6	500	0.037014	0.613663
4	2010_04	254	4	500	0.037013	0.677767
5	2010_05	279	6	500	0.031245	0.64261
6	2010_06	306	6	500	0.024996	0.628001
7	2010_07	301	4	500	0.023169	0.689071
8	2010_08	299	6	500	0.023592	0.649491
9	2010_09	275	2	500	0.023059	0.602698
10	2010_10	311	6	500	0.025625	0.726276
11	2010_11	307	6	500	0.0308	0.658046
12	2010_12	249	2	500	0.044703	0.464744

7. Random Forest model generation outputs for weekly layers in 2010 over the Midwest region of interest.

The **Layer** column shows the week of the year based on the first Julian day of each given week. The **No.Points** column shows the number of points used as training data in the entire sampling space in every week, this number corresponds with the 70% of the available sample points that were defined for training the model. The **Best_Mtry** column shows the number of covariates defined in the model in relation to the number of predictors used as input (6 predictors for this study: latitude, longitude, elevation, aspect, slope and topographic wetness index). The **Trees** column shows the maximum number of trees allows in the model generation. The **rmse** and **Cor** columns show the best results of the root minimum square error and the correlation respectively, derived from a cross validation process.

	Layer	No.Points	Best_Mtry	Trees	rmse	Cor
1	2010_001	259	2	500	0.035672	0.812821
2	2010_008	269	6	500	0.03306	0.821776
3	2010_015	309	4	500	0.031652	0.880383
4	2010_022	308	4	500	0.032526	0.862134
5	2010_029	289	2	500	0.043124	0.749933
6	2010_036	309	6	500	0.031207	0.755711
7	2010_043	311	4	500	0.031438	0.818046
8	2010_050	312	2	500	0.030511	0.827833
9	2010_057	314	2	500	0.031575	0.809027
10	2010_064	313	2	500	0.0326	0.824501
11	2010_071	314	4	500	0.031257	0.824936
12	2010_078	314	2	500	0.032426	0.815981
13	2010_085	313	4	500	0.030963	0.880842
14	2010_092	313	2	500	0.03182	0.91929
15	2010_099	313	4	500	0.027513	0.917342

16	2010_106	314	4	500	0.031687	0.649904
17	2010_113	313	4	500	0.03021	0.73991
18	2010_120	313	4	500	0.030687	0.856312
19	2010_127	314	4	500	0.030894	0.839125
20	2010_134	314	4	500	0.030755	0.750848
21	2010_141	313	2	500	0.030384	0.755212
22	2010_148	313	4	500	0.030674	0.776
23	2010_155	314	4	500	0.029479	0.844907
24	2010_162	314	4	500	0.031882	0.712714
25	2010_169	314	4	500	0.029571	0.814024
26	2010_176	313	4	500	0.02988	0.753661
27	2010_183	314	4	500	0.031329	0.760456
28	2010_190	313	4	500	0.030539	0.711995
29	2010_197	314	4	500	0.031376	0.763778
30	2010_204	313	2	500	0.03181	0.642023
31	2010_211	314	2	500	0.02905	0.675296
32	2010_218	313	2	500	0.030918	0.60397
33	2010_225	314	6	500	0.032493	0.621609
34	2010_232	314	6	500	0.031984	0.674627
35	2010_239	313	4	500	0.030664	0.715705
36	2010_246	313	4	500	0.032339	0.79463
37	2010_253	314	4	500	0.031788	0.831324
38	2010_260	313	6	500	0.031677	0.777231
39	2010_267	313	4	500	0.029879	0.756672
40	2010_274	314	4	500	0.029563	0.685711
41	2010_281	313	4	500	0.029012	0.786787
42	2010_288	314	2	500	0.028851	0.825059
43	2010_295	313	2	500	0.032469	0.536958
44	2010_302	313	6	500	0.029008	0.770491
45	2010_309	314	4	500	0.029448	0.781915
46	2010_316	313	4	500	0.032048	0.682287
47	2010_323	313	6	500	0.032005	0.76434
48	2010_330	310	2	500	0.030399	0.844596
49	2010_337	308	4	500	0.032859	0.805868
50	2010_344	298	2	500	0.031482	0.78782
51	2010_351	269	2	500	0.035504	0.706698
52	2010_358	235	2	500	0.035822	0.805875

8. **Random Forest model generation outputs for weekly layers in 2010 over the West region of interest.**

The **Layer** column shows the week of the year based on the first Julian day of each given week. The **No.Points** column shows the number of points used as training data in the entire sampling space in every week, this number corresponds with the 70% of the available sample points that were defined for training the model. The **Best_Mtry** column shows the number of covariates defined in the model in relation to the number of predictors used as input (6 predictors for this study: latitude, longitude, elevation, aspect, slope and topographic wetness index). The **Trees** column shows the maximum number of trees allows in the model generation. The **rmse** and **Cor** columns show the best results of the root minimum square error and the correlation respectively, derived from a cross validation process.

	Layer	No.Points	Best_Mtry	Trees	rmse	Cor
1	2010_001	82	4	500	0.04054	0.75674
2	2010_008	92	2	500	0.042483	0.652301
3	2010_015	134	4	500	0.041288	0.579426
4	2010_022	110	4	500	0.035949	0.773238
5	2010_029	79	4	500	0.040783	0.651056
6	2010_036	163	6	500	0.039855	0.595294
7	2010_043	112	4	500	0.040229	0.583188
8	2010_050	125	6	500	0.034191	0.695344
9	2010_057	118	4	500	0.035009	0.83284
10	2010_064	144	6	500	0.033577	0.666503
11	2010_071	116	4	500	0.041735	0.759273
12	2010_078	98	6	500	0.037456	0.740395
13	2010_085	112	6	500	0.035803	0.639456
14	2010_092	136	6	500	0.036887	0.795528
15	2010_099	139	2	500	0.037609	0.716505
16	2010_106	213	6	500	0.032693	0.814863
17	2010_113	245	6	500	0.034536	0.745326
18	2010_120	185	6	500	0.035367	0.635394
19	2010_127	212	6	500	0.034217	0.715174
20	2010_134	271	6	500	0.027533	0.692708
21	2010_141	232	6	500	0.034268	0.747985
22	2010_148	263	6	500	0.028933	0.757469
23	2010_155	282	4	500	0.033341	0.609875
24	2010_162	277	6	500	0.028214	0.664735
25	2010_169	183	2	500	0.030446	0.420546
26	2010_176	224	2	500	0.02964	0.487509
27	2010_183	213	2	500	0.027293	0.504962
28	2010_190	228	2	500	0.028027	0.412982
29	2010_197	273	4	500	0.026891	0.459981
30	2010_204	266	6	500	0.02732	0.757963

31	2010_211	250	4	500	0.025885	0.661996
32	2010_218	226	6	500	0.025321	0.712892
33	2010_225	230	4	500	0.028122	0.625634
34	2010_232	235	6	500	0.024251	0.655923
35	2010_239	283	6	500	0.024632	0.557175
36	2010_246	244	4	500	0.027099	0.68102
37	2010_253	232	4	500	0.02417	0.583953
38	2010_260	233	6	500	0.026441	0.437835
39	2010_267	225	4	500	0.027502	0.510778
40	2010_274	310	6	500	0.028572	0.692174
41	2010_281	308	6	500	0.027651	0.794839
42	2010_288	310	4	500	0.030202	0.685371
43	2010_295	310	2	500	0.031837	0.563831
44	2010_302	295	2	500	0.032907	0.640327
45	2010_309	296	6	500	0.033913	0.576586
46	2010_316	243	4	500	0.031364	0.676599
47	2010_323	203	4	500	0.036792	0.781788
48	2010_330	161	6	500	0.040691	0.685306
49	2010_337	179	4	500	0.047305	0.628554
50	2010_344	171	6	500	0.02946	0.825084
51	2010_351	207	4	500	0.037134	0.439452
52	2010_358	168	4	500	0.041899	0.563108