

Supplementary materials of the paper “Kernel-based estimation of individual location densities from smartphone data”

Earthquake network data

Figure 1 shows some relevant statistics about the signals when they are aggregated at the smartphone user level. Panels (a-f) show the distribution of the number of signals per smartphone, of the mean sampling interval, of the mean standard deviation, of the mean Euclidean distance between signals, of the maximum sampling interval and of the maximum standard deviation, respectively. The average sampling interval is around 62 minutes while the average Euclidean distance is around 580 meters. The maximum interval between two signals from the same smartphone has an average of around seven days but it can be as high as one or two months, while σ_t can be as high as 40 kilometres. These summaries highlight that the interval between two signals can be very high as well as the sampling interval variability can be high both within and between smartphones. This calls for a location estimation approach which is able to exploit historical information on the smartphone location rather than a tracking approach based on the nearest observations.

Trimming

In order to understand the implications of trimming the components in (2.2), we consider one simulated data set of Section 4 of the paper and we estimate the parameters under $\delta = \{2, 5, 10, 20, 50, 100, 150, 200, 250, 500, 1000\}$ (see equation 3.5 of the paper). A comparison is then made with respect to the non-trimmed estimation and it is based on the following mean absolute

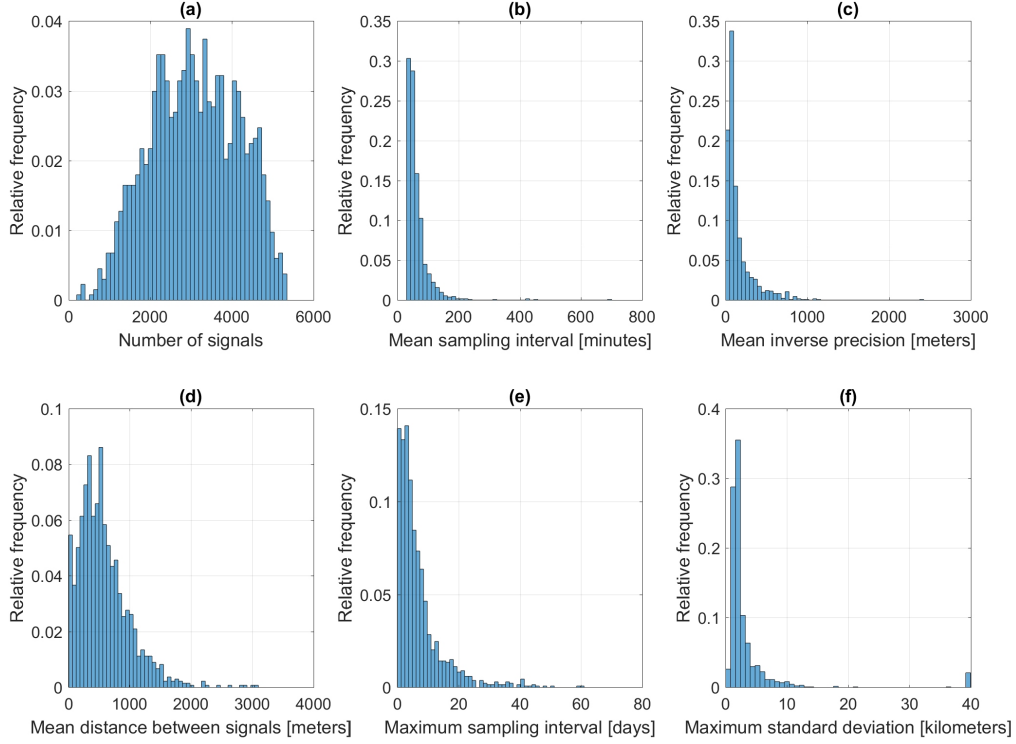


Figure 1: Histograms summarizing the signals collected by the Earthquake Network project and aggregated for the 1,188 smartphones. Panel (a): number of signals sent to the server by each smartphone in the period January, 1st - April, 30th 2017; panel (b): mean sampling interval; panel (c): mean standard deviation on the observed locations; panel (d): mean distance between two consecutive signals; panel (e): maximum sampling interval (gap) between two consecutive signals; panel (f): maximum standard deviation on the observed locations.

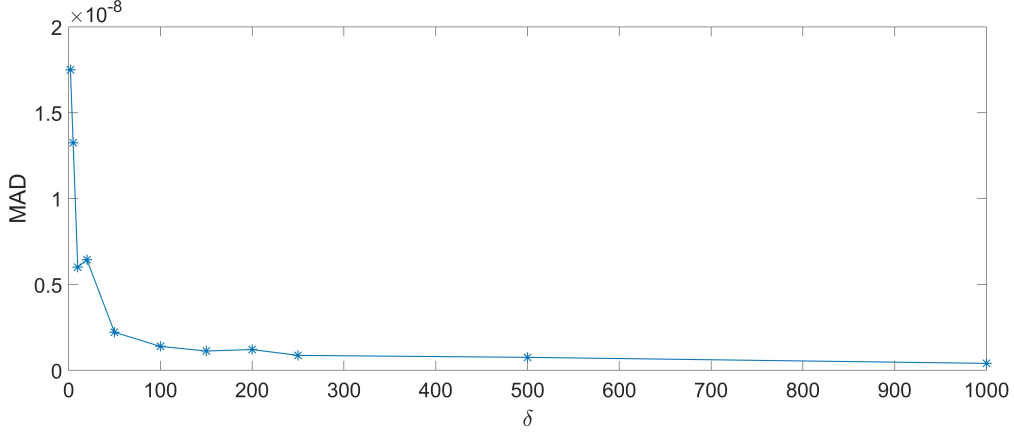


Figure 2: Mean absolute difference (MAD) between the estimated location density when the trimming is not enabled and the analogous trimmed estimate, as a function of threshold δ .

difference (MAD)

$$MAD = \frac{1}{N} \sum_{t=1}^T \sum_{\mathbf{s}^* \in \mathcal{D}^*} \left| \hat{f}_{t|T}(\mathbf{s}^*; \hat{\boldsymbol{\theta}}) - \tilde{f}_{t|T}(\mathbf{s}^*; \hat{\boldsymbol{\theta}}, \delta) \right|, \quad (1)$$

where \hat{f} and \tilde{f} are the non-trimmed and the trimmed estimates, respectively, \mathcal{D}^* is the discretised spatial domain, \mathbf{s}^* is the centre of the pixel and N is the total number of sum terms.

The comparison results are depicted in Figure 2. As expected, the difference decreases as δ increases, i.e., the trimming effect attenuates as the number of components increases. With a threshold larger than 250, the MAD drop is reduced, suggesting that 250 might be a fair choice for the simulated data set.

For one of the 1,188 smartphones considered in Section 5 of the paper, Figure 3 shows the variation along time of the number mixture components when $\delta = 250$. Note that the number of components ranges from 80 to 350 out of roughly 3,700 total locations available for such smartphone.

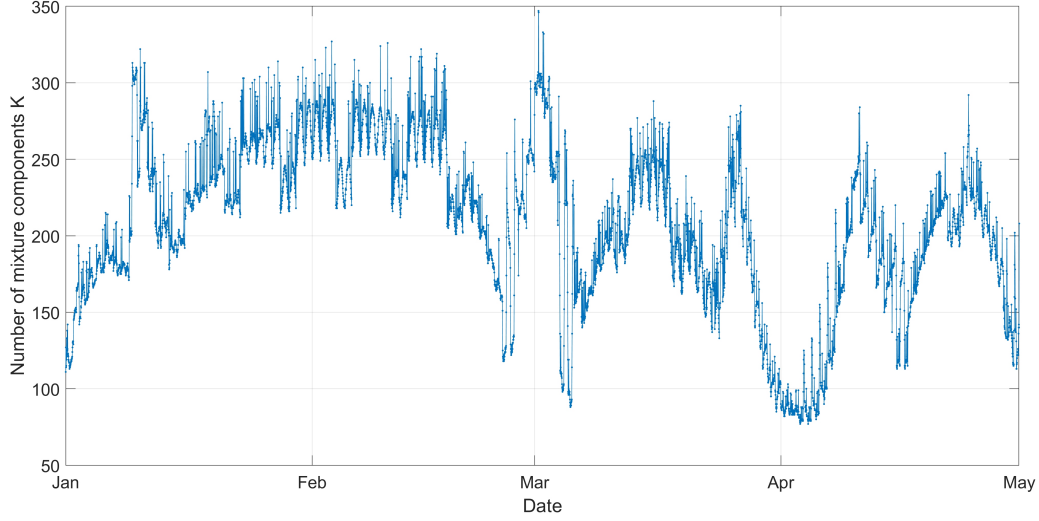


Figure 3: The number of mixture components used to estimate the density location with trimming enabled for a given smartphone.

Simulation study

A simulation study is carried out to clarify the weighting scheme in (3.4) and to show the capability to recover model parameters using the LCV estimation procedure. The ingredients of the simulation are the sampling times and the smartphone locations together with their precisions. We simulate location data from January 1st, 2017 to March 1st, 2017 (60 days) for $n = 100$ smartphones. The sampling time of each smartphone i ($i = 1, \dots, n$) is given by $t_{i,j+1} = t_{i,j} + \Delta_{i,j} + \lambda$ ($j = 1, \dots, m_i$) where $\Delta_{i,j}$ is sampled from the exponential distribution $\text{Exp}(\lambda)$, with $\lambda = \{24, 4\}$.

The initial observed location is set as $\mathbf{s}_{i,0} = (0, 0)'$ with precision $1/\sigma_{i,0}^2$, where $\sigma_{i,0}$ is sampled from the generalized extreme value distribution $\text{GEV}(0.5, 10, 20)$, the parameters of which are chosen such that the values of $\sigma_{i,j}$ mimic those observed in the real case. We consider a full factorial design coming from the following values: $\phi_1 = \{1, 10\}$, $\phi_2 = \{0.05, 0.3\}$, $\phi_3 = \{0.5, 3\}$ and $\alpha = \{0.5, 10\}$. In other words, we simulate 16 scenarios for each value of λ that leads to different number of locations with expected values 1416 and 236, respectively.

Figure 4 shows the unnormalized weights in (3.3) associated with $\alpha =$

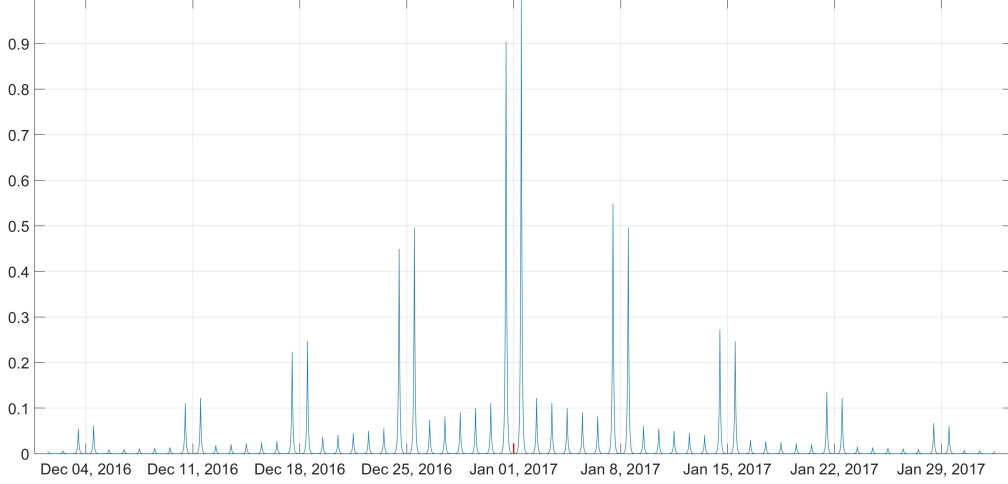


Figure 4: Unnormalized weights $u(t, t'; \phi)$ when t is January 1st, 2017 12 PM and $\phi = (10, 0.05, 0.5)$.

$1, \phi_1 = 10, \phi_2 = 0.05, \phi_3 = 0.5$. In particular, the plot represents $u(t, t'; \phi)$ as a function of time, when t is January 1st, 2017 at 12 PM. The maxima correspond to the same time in previous days while the weights drop to zero when moving far from 12 PM. Moreover, since t is Sunday, weights are higher for times belonging to the same weekend and to all the previous and future weekends, while working days have a lower weight. Weights decay exponentially over time and after four weeks they approach zero.

At any time, locations are simulated as follows. For each time $t_{i,j}$, the true smartphone location $\tilde{\mathbf{s}}_{i,j}$ is sampled sequentially from (3.2) given the information set up to time $t_{i,j-1}$. The precision at time $t_{i,j}$ is obtained sampling $\sigma_{i,j}$ from the GEV distribution mentioned above. Finally, the observed smartphone location $\mathbf{s}_{i,j}$ is sampled from $\mathcal{N}_2(\tilde{\mathbf{s}}_{i,j}, \Sigma_{i,j})$, with $\Sigma_{i,j} = \sigma_{i,j}^2 \mathbf{I}_2$. Note that $\tilde{\mathbf{s}}_{i,j}$ is the true smartphone location while $\mathbf{s}_{i,j}$ simulates the location provided by the smartphone and sent to the server. The information available at time $t_{i,j}$ is thus $\varphi(\mathbf{s}; \mathbf{s}_{i,j}, \Sigma_{i,j})$, the normal density centred on $\mathbf{s}_{i,j}$ with covariance $\Sigma_{i,j}$.

As an example, Figure 5 shows the simulated locations for a single smartphone; \mathbf{s}_j is the centre of the disk while σ_j the radius. In Figure 5, smartphone locations appear clustered in space since the smartphone tends to be at the same location at the same time within the day. Moreover, locations far from

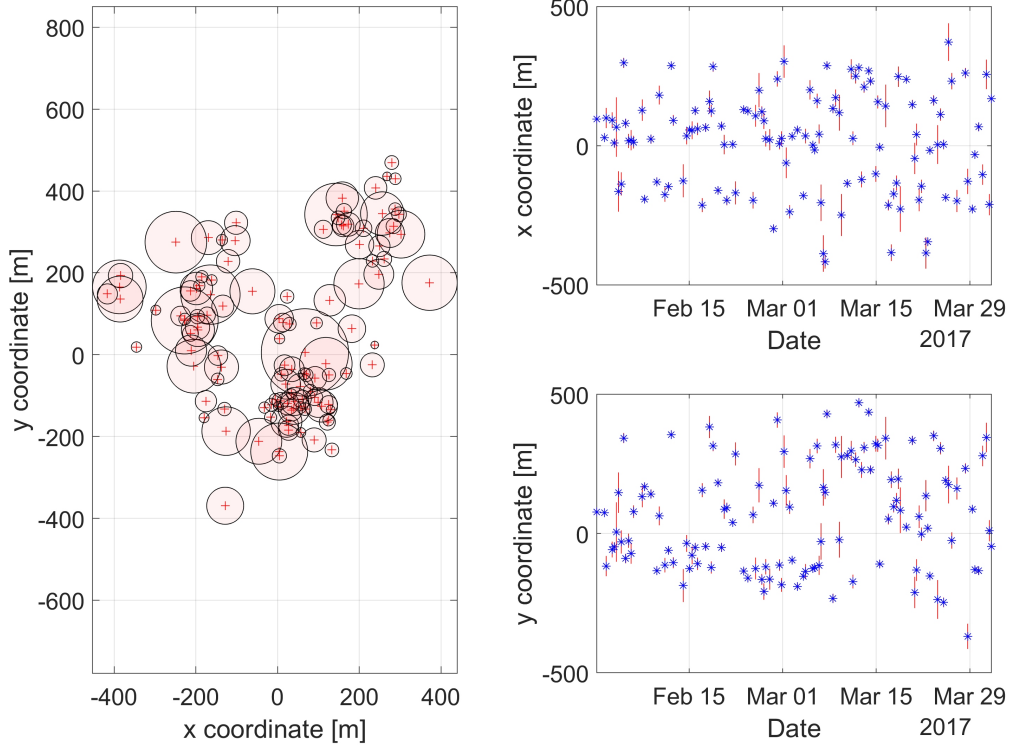


Figure 5: Spatial (left panel) and temporal (right panels) distribution of the simulated locations for a single smartphone. Markers are the observed locations \mathbf{s}_j while the radius of the disks and the vertical segments are the standard deviations σ_j .

the cluster centres tend to have lower precisions (high disk radius). The true smartphone location is likely near to the cluster centre but the lower precision implies that the observed location is not.

For each simulated smartphone, the parameter vector $\boldsymbol{\theta}$ is estimated by maximising the LCV function as detailed in Section 3 without any trimming. Initial values are randomly sampled from uniform distributions centred on the true value of $\boldsymbol{\theta}$ and with range equal to the element itself. Therefore, we also test the robustness of the maximisation algorithm in finding the maximum starting far from the true values of the model parameters.

Table 1 reports the estimation results in terms of mean, 5th and 95th percentile of the empirical distribution obtained for each parameter of $\boldsymbol{\theta}$ over

the 100 simulated smartphones for all scenarios. Results suggest that the model is identifiable and that the estimator of $\boldsymbol{\theta}$ is not biased for most of the cases. Looking at the 95th percentile, however, we notice that parameter ϕ_3 is not always identifiable. This may happen in two cases: when the user does not change her/his behaviour during the weekend or when locations collected during weekends are small in number. In this case, ϕ_3 tends to be a large number and the third exponential term in equation (3.4) of the paper is nearly 1, namely there is no weekend effect.

Within the simulation setting, we also look at the capability of recovering the true smartphone location \mathbf{s}_t from the estimated density $f_{t|T}$ when employing a point estimate, say the average. Table 2 offers a comparison between our time-varying estimator (3.2) with its time-invariant version in (3.1) and the DLM in (6.1) - (6.2). A cross-validation study based on 20% of held-out locations is performed and the comparison is provided in terms of average euclidean distance between the simulated locations and the estimated ones. Our estimator outperforms the other two approaches in all scenarios, with higher predictive performance when locations are less persistent in time, i.e., ϕ_1 is small.

Estimation results

The histograms in Figure 6 displays the estimated $\hat{\alpha}$ values and the estimated $\hat{\phi}$ values for the 1,188 smartphones. Recall that a high value of $\hat{\alpha}$ implies a large uncertainty on the smartphone location at time t while the $\hat{\phi}$ values describe the temporal persistence.

	True				α			ϕ_1			ϕ_2			ϕ_3		
λ	α	ϕ_1	ϕ_2	ϕ_3	5th	Mean	95th	5th	Mean	95th	5th	Mean	95th	5th	Mean	95th
4	0.5	1	0.05	0.5	0.00	0.14	0.82	0.71	1.16	1.71	0.03	0.05	0.08	0.32	0.82	4.05
4	10	1	0.05	0.5	4.53	7.52	11.20	0.84	1.28	1.66	0.04	0.06	0.08	0.37	0.79	2.61
4	0.5	1	0.05	3	0.00	0.07	0.93	0.73	1.11	1.66	0.04	0.05	0.07	1.20	3.01	117095.73
4	10	1	0.05	3	4.46	7.84	11.50	0.89	1.18	1.99	0.04	0.06	0.08	0.82	4.51	202453.76
4	0.5	1	0.3	0.5	0.00	0.03	0.97	0.61	0.88	1.41	0.14	0.31	3.73	0.20	0.64	4.57
4	10	1	0.3	0.5	3.89	6.89	13.68	0.63	1.01	1.63	0.17	0.33	4414.40	0.34	0.75	4.41
4	0.5	1	0.3	3	0.00	0.03	0.71	0.68	0.93	1.31	0.14	0.24	1.38	1.10	3.00	306841.66
4	10	1	0.3	3	3.98	6.81	10.65	0.74	1.12	1.56	0.15	0.32	658.85	0.86	15.18	189219.34
4	0.5	10	0.05	0.5	0.00	0.29	1.14	6.83	10.05	18.19	0.04	0.05	0.07	0.26	0.45	1.18
4	10	10	0.05	0.5	5.82	8.74	12.85	7.09	11.26	17.55	0.04	0.05	0.07	0.30	0.49	1.18
4	0.5	10	0.05	3	0.00	0.14	1.08	6.66	10.02	16.09	0.03	0.05	0.07	0.85	3.00	225343.79
4	10	10	0.05	3	3.90	7.93	13.68	6.64	10.05	17.57	0.04	0.05	0.07	1.05	4.45	93837.87
4	0.5	10	0.3	0.5	0.00	0.29	1.35	6.46	10.00	17.02	0.14	0.30	1.42	0.00	0.42	1.33
4	10	10	0.3	0.5	4.36	8.35	13.28	6.13	10.61	18.70	0.14	0.32	2.09	0.00	0.43	1.57
4	0.5	10	0.3	3	0.00	0.20	1.09	6.65	9.98	14.78	0.14	0.26	1.61	0.91	2.99	156938.27
4	10	10	0.3	3	3.36	7.35	14.71	6.18	10.69	19.30	0.15	0.26	1.18	0.81	5.91	177180.75
24	0.5	1	0.05	0.5	0.00	0.16	0.38	0.78	0.93	1.07	0.04	0.05	0.05	0.41	0.51	0.74
24	10	1	0.05	0.5	5.91	7.38	9.04	0.83	0.94	1.09	0.04	0.05	0.05	0.41	0.54	0.74
24	0.5	1	0.05	3	0.00	0.17	0.41	0.84	0.94	1.11	0.04	0.04	0.05	1.41	3.00	14668.25
24	10	1	0.05	3	5.96	7.66	8.95	0.87	0.98	1.12	0.04	0.05	0.05	1.49	3.54	10200.77
24	0.5	1	0.3	0.5	0.00	0.05	0.44	0.75	0.90	1.08	0.19	0.26	0.40	0.28	0.50	0.79
24	10	1	0.3	0.5	5.19	6.72	8.28	0.78	0.94	1.08	0.21	0.27	0.39	0.34	0.54	0.83
24	0.5	1	0.3	3	0.00	0.09	0.39	0.80	0.92	1.10	0.19	0.26	0.41	1.34	3.00	9345.43
24	10	1	0.3	3	5.50	6.88	8.70	0.82	0.99	1.18	0.21	0.26	0.43	1.48	4.61	32494.09
24	0.5	10	0.05	0.5	0.00	0.27	0.59	8.21	10.00	13.52	0.04	0.05	0.06	0.35	0.46	0.65
24	10	10	0.05	0.5	5.84	8.18	11.02	8.65	10.14	12.70	0.04	0.05	0.06	0.38	0.46	0.61
24	0.5	10	0.05	3	0.00	0.25	0.60	8.43	10.03	12.73	0.04	0.05	0.06	1.31	3.02	23795.08
24	10	10	0.05	3	5.91	7.64	9.41	8.01	10.15	12.27	0.04	0.05	0.06	1.34	3.51	12647.59
24	0.5	10	0.3	0.5	0.00	0.16	0.58	7.24	9.99	12.99	0.18	0.28	0.62	0.31	0.44	0.69
24	10	10	0.3	0.5	5.19	7.25	9.80	7.83	10.45	14.14	0.20	0.29	0.54	0.31	0.48	0.67
24	0.5	10	0.3	3	0.00	0.17	0.88	7.17	9.93	12.13	0.18	0.26	0.64	1.05	2.99	1369.67
24	10	10	0.3	3	4.68	7.35	9.75	7.69	9.83	13.80	0.20	0.30	0.60	1.22	3.18	28659.94

Table 1: Parameter estimates for 100⁸ simulated datasets over all scenarios.

	True				Distance			Relative difference	
λ	α	ϕ_1	ϕ_2	ϕ_3	DLM	Invariant	MIX	MIX vs Invariant	MIX vs DLM
4	0.5	1	0.05	0.5	335.10	496.76	123.56	-75%	-63%
4	10	1	0.05	0.5	722.34	844.19	276.29	-67%	-62%
4	0.5	1	0.05	3	315.40	349.64	110.98	-68%	-65%
4	10	1	0.05	3	809.44	921.50	260.77	-72%	-68%
4	0.5	1	0.3	0.5	288.25	313.43	103.91	-67%	-64%
4	10	1	0.3	0.5	632.37	735.43	220.56	-70%	-65%
4	0.5	1	0.3	3	376.52	436.20	119.08	-73%	-68%
4	10	1	0.3	3	654.58	771.81	258.16	-67%	-61%
4	0.5	10	0.05	0.5	249.74	322.49	196.80	-39%	-21%
4	10	10	0.05	0.5	534.22	648.08	407.24	-37%	-24%
4	0.5	10	0.05	3	236.82	294.22	195.28	-34%	-18%
4	10	10	0.05	3	527.56	669.67	402.30	-40%	-24%
4	0.5	10	0.3	0.5	232.98	280.64	223.48	-20%	-4%
4	10	10	0.3	0.5	531.32	745.35	486.95	-35%	-8%
4	0.5	10	0.3	3	235.81	285.74	231.62	-19%	-2%
4	10	10	0.3	3	543.41	686.81	522.96	-24%	-4%
24	0.5	1	0.05	0.5	595.17	575.08	257.59	-55%	-57%
24	10	1	0.05	0.5	1225.81	1222.92	535.75	-56%	-56%
24	0.5	1	0.05	3	598.88	606.26	295.79	-51%	-51%
24	10	1	0.05	3	1473.85	1615.60	682.69	-58%	-54%
24	0.5	1	0.3	0.5	547.16	483.35	244.46	-49%	-55%
24	10	1	0.3	0.5	1347.65	1260.93	622.12	-51%	-54%
24	0.5	1	0.3	3	580.82	525.04	319.26	-39%	-45%
24	10	1	0.3	3	1503.48	1198.59	860.58	-28%	-43%
24	0.5	10	0.05	0.5	292.41	305.43	270.14	-12%	-8%
24	10	10	0.05	0.5	710.69	827.22	635.11	-23%	-11%
24	0.5	10	0.05	3	291.56	311.33	274.96	-12%	-6%
24	10	10	0.05	3	686.70	825.20	626.40	-24%	-9%
24	0.5	10	0.3	0.5	327.58	339.23	322.33	-5%	-2%
24	10	10	0.3	0.5	748.99	892.92	742.28	-17%	-1%
24	0.5	10	0.3	3	309.22	317.65	307.66	-3%	-1%
24	10	10	0.3	3	712.37	816.53	679.53	-17%	-5%

Table 2: Model comparison over all scenarios; dynamic linear model (DLM), time-invariant density estimator (Invariant) and time-varying location density (MIX).

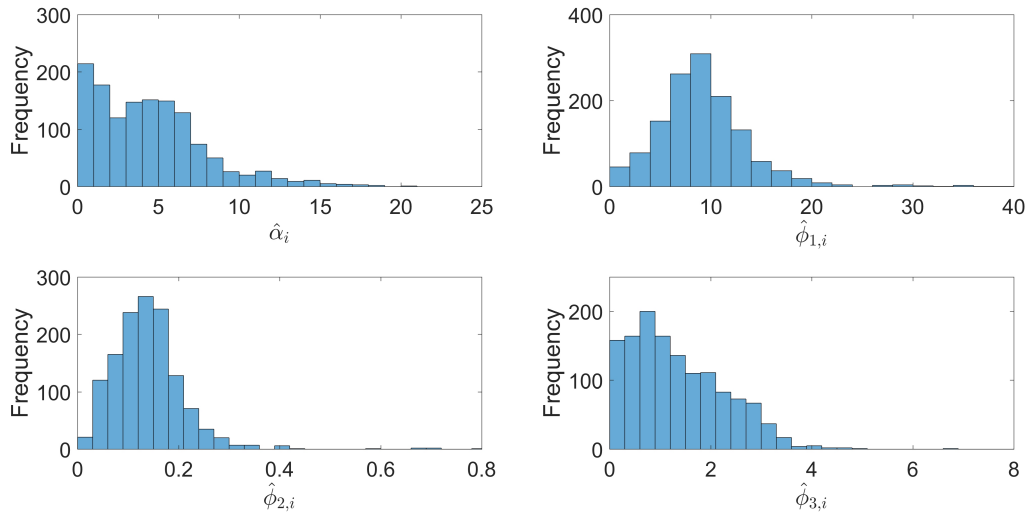


Figure 6: Distributions of the estimated parameters for all 1,188 smartphones.