Supplementary Material to Reproducibility Report for Reproducibility Challenge 2021

Anonymous Author(s)
Affiliation
Address
`email`

Level-wise Model Performance

In Tables 1 through 5 we report the CRPS scores of the models for the time series aggregated at a certain level in a dataset (1. level being the **root** of a hierarchy). As already mentioned in our **Reproducibility Report for Reproducibility Challenge 2021**, the scores represent the average over 5 runs. State-of-the-art methods except for `PERMBU-MinT` produced the same results over all runs, thus no uncertainty is being reported.

For reference, in the second column of each table we also include the mean CRPS score of a model on a particular dataset. The scores in **boldface** represent the best performance in a column. And, the scores in $\boxed{\text{boxes}}$ denote the lowest CRPS score in a column, but taking into consideration all models except `HierE2E` and its two variants.

Methods	Mean	1. level	2. level	3. level	4. level
`ARIMA-NaiveBU`	0.0453	0.0437	0.0441	0.0447	0.0489
`ETS-NaiveBU`	0.0432	0.0416	0.0418	0.0421	0.0471
`ARIMA-MinT-shr`	0.0467	0.0454	0.0455	0.0459	0.0499
`ARIMA-MinT-ols`	0.0463	0.0448	0.0450	0.0455	0.0499
`ETS-MinT-shr`	0.0455	0.0440	0.0442	0.0444	0.0492
`ETS-MinT-ols`	0.0459	0.0445	0.0447	0.0448	0.0495
`ARIMA-ERM`	0.0399	$\boxed{0.0365}$	$\boxed{0.0379}$	0.0391	0.0459
`ETS-ERM`	0.0456	0.0409	0.0437	0.0452	0.0525
`PERMBU-MinT`	$\boxed{0.0393 \pm 0.0003}$	0.0406 ± 0.0004	0.0388 ± 0.0003	$\boxed{0.0382 \pm 0.0002}$	$\boxed{0.0396 \pm 0.0003}$
`HierE2E`	**0.0335** ± **0.0072**	**0.0302** ± **0.0093**	**0.0342** ± **0.0071**	**0.0335** ± **0.0066**	**0.0361** ± **0.0058**
`DeepVAR`	0.0367 ± 0.0055	0.0342 ± 0.0050	0.0362 ± 0.0059	0.0362 ± 0.0056	0.0403 ± 0.0067
`DeepVAR+`	0.0457 ± 0.0130	0.0445 ± 0.0160	0.0461 ± 0.0130	0.0456 ± 0.0125	0.0466 ± 0.0106

Table 1: Here we present the level-wise CRPS scores of the models on the dataset `Labour`. As it can be observed, `HierE2E` outperforms the other models, overall and at hierarchical level. For the classical machine learning methods, it is to be noted that `PERMBU-MinT` and `ARIMA-ERM` have competitive results to the ones of the neural approaches, each being the best non-neural technique at two levels, and `PERMBU-MinT` also obtaining the lowest mean score in this group of techniques.

Methods	Mean	1. level	2. level	3. level	4. level
ARIMA-NaiveBU	0.1138	0.0588	0.0945	0.1366	0.1653
ETS-NaiveBU	0.1008	0.0545	0.0809	0.1194	0.1483
ARIMA-MinT-shr	0.1171	0.0625	0.0989	0.1395	0.1677
ARIMA-MinT-ols	0.1195	0.0619	0.1018	0.1419	0.1723
ETS-MinT-shr	0.1013	0.0592	0.0793	0.1202	0.1467
ETS-MinT-ols	0.1002	0.0597	0.0749	0.1201	0.1462
ARIMA-ERM	0.5885	0.2196	0.3903	0.8120	0.9322
ETS-ERM	2.3742	1.4383	1.9934	2.8479	3.2173
PERMBU-MinT	**0.0763 ± 0.0003**	$\boxed{\mathbf{0.0464 \pm 0.0017}}$	$\boxed{\mathbf{0.0592 \pm 0.0008}}$	$\boxed{\mathbf{0.0899 \pm 0.0011}}$	$\boxed{\mathbf{0.1097 \pm 0.0009}}$
HierE2E	0.0916 ± 0.0091	0.0510 ± 0.0099	0.0765 ± 0.0113	0.1104 ± 0.0080	0.1286 ± 0.0079
DeepVAR	0.0953 ± 0.0062	0.0531 ± 0.0120	0.0827 ± 0.0091	0.1120 ± 0.0086	0.1333 ± 0.0062
DeepVAR+	0.0956 ± 0.0180	0.0509 ± 0.0190	0.0776 ± 0.0216	0.1148 ± 0.0180	0.1390 ± 0.0152

Table 2: Here we present the level-wise CRPS scores of the models on the dataset Tourism. It is clear the model PERMBU-MinT would be the optimal choice for this data, because it makes the best predictions at every level of the hierarchy. It is also worth pointing out that the models ARIMA-ERM and ETS-ERM perform worse on this dataset than the other modelling techniques.

Methods	Mean	1. level	2. level (geo.)	3. level (geo.)	4. level (geo.)	2. level (trav.)	3. level (trav.)	4. level (trav.)	5. level (trav.)
ARIMA-NaiveBU	0.1752	0.0827	0.1035	0.1586	0.2131	0.1003	0.1567	0.2489	0.3379
ETS-NaiveBU	0.1690	0.0802	0.0989	0.1561	0.2058	0.0927	0.1484	$\boxed{0.2408}$	$\boxed{0.3291}$
ARIMA-MinT-shr	$\boxed{0.1615}$	0.0443	$\boxed{0.0826}$	$\boxed{0.1439}$	0.2042	$\boxed{\mathbf{0.0834}}$	0.1485	0.2440	0.3413
ARIMA-MinT-ols	0.1731	$\boxed{\mathbf{0.0394}}$	0.0830	0.1501	0.2169	0.1056	0.1646	0.2610	0.3643
ETS-MinT-shr	0.1627	0.0505	0.0902	0.1501	$\boxed{0.2024}$	0.0890	$\boxed{0.1439}$	0.2415	0.3343
ETS-MinT-ols	0.1668	0.0484	0.0897	0.1542	0.2102	0.0891	0.1455	0.2499	0.3473
ARIMA-ERM	0.5668	0.2577	0.3791	0.4974	0.6380	0.3660	0.5402	0.8013	1.0551
ETS-ERM	0.5080	0.1161	0.3231	0.4684	0.6143	0.2622	0.4853	0.7741	1.0209
PERMBU-MinT	-	-	-	-	-	-	-	-	-
HierE2E	0.1688 ± 0.0040	0.0959 ± 0.0105	0.1161 ± 0.0063	0.1503 ± 0.0053	0.1901 ± 0.0045	0.1209 ± 0.0039	0.1619 ± 0.0044	0.2242 ± 0.0044	0.2913 ± 0.0053
DeepVAR	**0.1394 ± 0.0021**	0.0634 ± 0.0050	**0.0814 ± 0.0029**	**0.1216 ± 0.0030**	**0.1629 ± 0.0017**	0.0891 ± 0.0087	**0.1302 ± 0.0040**	**0.1979 ± 0.0012**	**0.2684 ± 0.0026**
DeepVAR+	0.1979 ± 0.0294	0.1234 ± 0.0430	0.1417 ± 0.0351	0.1775 ± 0.0304	0.2180 ± 0.0263	0.1464 ± 0.0331	0.1895 ± 0.0259	0.2556 ± 0.0234	0.3314 ± 0.0245

Table 3: Here we present the level-wise CRPS scores of the models on the dataset Tourism-L. As it can be observed, DeepVAR is the best performing model at 6 levels as well as overall, but ARIMA-MinT-ols and ARIMA-MinT-shr also achieve the best result at one level each. If we analyze only the models other than HierE2E, DeepVAR and DeepVAR+, we come to the conclusion there is no one most dominant model, but as many as four different models manage to have the lowest CRPS score at some level in the hierarchy. Expectedly, as Tourism-L represents an extension of Tourism, the two models using the ERM reconciliation technique again show worse performance than the competitors.

Methods	Mean	1. level	2. level	3. level	4. level
ARIMA-NaiveBU	0.0753	0.0364	0.0364	0.0453	0.1832
ETS-NaiveBU	0.0665	0.0128	0.0128	0.0351	0.2053
ARIMA-MinT-shr	0.0775	0.0467	0.0467	0.0467	0.1701
ARIMA-MinT-ols	0.1123	0.0853	0.0853	0.0853	0.1934
ETS-MinT-shr	0.0963	0.0601	0.0601	0.0601	0.2050
ETS-MinT-ols	0.1110	0.0765	0.0765	0.0765	0.2145
ARIMA-ERM	$\boxed{0.0466}$	$\boxed{\mathbf{0.0089}}$	$\boxed{\mathbf{0.0113}}$	$\boxed{0.0254}$	$\boxed{0.1408}$
ETS-ERM	0.1027	0.0828	0.0828	0.0828	0.1624
PERMBU-MinT	0.0679 ± 0.0053	0.0346 ± 0.0072	0.0354 ± 0.0058	0.0419 ± 0.0044	0.1598 ± 0.0042
HierE2E	0.0359 ± 0.0127	0.0166 ± 0.0170	0.0178 ± 0.0159	**0.0186 ± 0.0154**	0.0905 ± 0.0061
DeepVAR	**0.0334 ± 0.0036**	0.0131 ± 0.0058	0.0174 ± 0.0121	0.0198 ± 0.0086	**0.0835 ± 0.0027**
DeepVAR+	0.0366 ± 0.0088	0.0130 ± 0.0081	0.0158 ± 0.0080	0.0209 ± 0.0124	0.0969 ± 0.0096

Table 4: Here we present the level-wise CRPS scores of the models on the dataset Traffic. ARIMA-ERM is the model with the best CRPS score at most levels (2), whereas HierE2E and DeepVAR outperform the competition at only a single level each. However, what is pecular to these results is that ARIMA-ERM trades off the favorable accuracy in the upper levels for less favorable one towards the bottom of the hierarchy. Consequently, its score is higher than the scores of HierE2E and DeepVAR at the 3. and 4. level in the hierarchy.

Methods	Mean	1. level	2. level	3. level	4. level	5. level
ARIMA-NaiveBU	0.3776	0.1904	0.2797	0.4118	0.4124	0.5936
ETS-NaiveBU	0.4673	0.341	0.3863	0.4631	0.5051	0.641
ARIMA-MinT-shr	0.2466	0.08	0.1382	0.2559	0.2953	0.4638
ARIMA-MinT-ols	0.2782	0.1079	0.1743	0.2857	0.3253	0.4977
ETS-MinT-shr	0.3622	0.218	0.2666	0.3451	0.388	0.5936
ETS-MinT-ols	0.2702	$\boxed{\mathbf{0.0234}}$	0.1456	0.2616	0.3138	0.6065
ARIMA-ERM	0.2195	0.0776	0.1213	0.2325	0.2746	0.3913
ETS-ERM	0.2217	0.1558	0.1614	0.201	0.2399	0.3506
PERMBU-MinT	0.279 ± 0.0223	0.094 ± 0.0394	0.1599 ± 0.0248	0.2689 ± 0.0293	0.3056 ± 0.0305	0.5666 ± 0.0589
HierE2E	**0.1629 ± 0.0063**	0.0668 ± 0.0056	0.1184 ± 0.0062	**0.1536 ± 0.0082**	**0.1711 ± 0.0067**	**0.3047 ± 0.0076**
DeepVAR	0.2081 ± 0.0067	0.0751 ± 0.0153	0.1199 ± 0.0143	0.2238 ± 0.0074	0.2555 ± 0.0109	0.3663 ± 0.0047
DeepVAR+	0.2053 ± 0.0146	0.0523 ± 0.0158	**0.1053 ± 0.009**	0.2076 ± 0.0187	0.2567 ± 0.0205	0.4047 ± 0.0223

Table 5: Here we present the level-wise CRPS scores of the models on the dataset Wiki. The newly proposed model HierE2E is the optimal approach at the 3 bottom hierarchy levels, but also overall. ETS-MinT-ols has a lower CRPS score than the other models at the **root** of the hierarchy. Nevertheless, as we progress to the bottom hierarchy levels, we observe its performance becomes worse faster than the performance of any other model.