A Existing OQ methods from the quantification literature

For completeness, we introduce the OQ methods OQT[8] and ARC[9], which appear in our main experiments. Both methods address ordinality not via regularization, like we propose, but via recursive bipartitions of the codeframe.

A.1 Ordinal Quantification Tree (OQT)

The OQT algorithm trains a quantifier by arranging probabilistic binary classifiers (one for each possible bipartition of the ordered set of classes) into an *ordinal quantification tree* (OQT), which is conceptually similar to a hierarchical classifier. Two characteristic aspects of training an OQT are that (a) the loss function used for splitting a node is a quantification loss (and not a classification loss), e.g., the Kullback-Leibler Divergence, and (b) the splitting criterion is informed by the class order. Given a test document, one generates a posterior probability for each of the classes by having the document descend all branches of the trained tree. After this is done for all documents in the test sample, the probabilistic classify-and-count (PCC) multiclass (i.e., non-ordinal) quantification method is invoked in order to compute the final prevalence estimates.

The OQT method was only tested in the SemEval 2016 “Sentiment analysis in Twitter” shared task[10]. While OQT was the best performer in that subtask, its true value still has to be assessed, since the above-mentioned subtask evaluated participating algorithms on one test sample only. In our experiments, we have tested OQT in a much more robust way.

A.2 Adjusted Regress and Count (ARC)

The ARC algorithm is similar to OQT in that it trains a hierarchical classifier where the leaves of the tree are the classes, these leaves are ordered left-to-right, and each internal node partitions an ordered sequence of classes in two such subsequences. One difference between the two algorithms is the criterion used in order to decide where to split a given sequence of classes, which for OQT is based on a quantification loss (KLD), and for ARC is based on the principle of minimizing the imbalance (in terms of the number of training examples) of the two subsequences. A second difference is that, once the tree is trained and used to classify the test documents, OQT uses what is basically a PCC algorithm, while ARC uses the adjusted classify-and-count (ACC) multiclass quantification method.

[8] Da San Martino, G., Gao, W., Sebastiani, F.: Ordinal text quantification. In: SIGIR. pp. 937–940 (2016)

[9] Esuli, A.: ISTI-CNR at SemEval-2016 Task 4: Quantification on an ordinal scale. In: SEMEVAL. pp. 92–95 (2016)

[10] Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., Stoyanov, V.: SemEval-2016 Task 4: Sentiment analysis in Twitter. In: SEMEVAL. pp. 1–18 (2016)

Concerning the quality of ARC, the same considerations made for OQT apply, since ARC, like OQT, has only been tested in the Ordinal Quantification subtask of the SemEval 2016 “Sentiment analysis in Twitter” shared task; despite the fact that it worked well in that context, the experiments that we are presenting in our paper are more conclusive.

B Extended results

The following results complete the experiments we have shown in the main paper.

B.1 Performance in terms of RNOD

We have repeated all of our experiments in terms of the *Root Normalized Order-aware Divergence* (RNOD) evaluation measure, instead of NMD, as proposed by Sakai[11] and defined as

$$\mathrm{RNOD}(p, \hat{p}) = \left(\frac{\sum_{y_i \in \mathcal{Y}^*} \sum_{y_j \in \mathcal{Y}} d(y_j, y_i)(p(y_j) - \hat{p}(y_j))^2}{|\mathcal{Y}^*|(n-1)} \right)^{\frac{1}{2}} \tag{16}$$

where $\mathcal{Y}^* = \{y_i \in \mathcal{Y} | p(y_i) > 0\}$. Note that by adopting RNOD we are not simply replacing the evaluation measure, but also the criterion for model selection. That is to say, we have re-run all experiments, this time optimizing hyperparameters for RNOD in place of NMD.

From examining the RNOD results from Tab. 3, we may note that, while some methods change positions in the ranking, as compared to their ranks in terms of NMD, the general conclusions from the NMD evaluation in the main paper also hold in terms of RNOD.

We do not choose RNOD as the main evaluation function (and prefer NMD for the main paper instead) because we do not think RNOD is a satisfactory measure for OQ. The reason why we do not consider RNOD a satisfactory OQ measure is that, without (we think) reason, it penalizes more heavily mistakes (i.e., “transfers” of probability mass from a class to another) closer to the extremes of the codeframe. For instance, given $\mathcal{Y} = \{y_1, y_3, y_3, y_4, y_5\}$, assume $p = (0.2, 0.2, 0.2, 0.2, 0.2)$, and assume two predicted distributions $\hat{p}' = (0.2, 0.2, 0.3, 0.1, 0.2)$ and $\hat{p}'' = (0.2, 0.2, 0.2, 0.3, 0.1)$. The two predicted distributions make essentially the same mistake, i.e., erroneously “transfer” a probability mass of 0.1 from a class y_i to a class $y_{(i-1)}$, the difference being that in $\hat{p}'$ it is the case that $i = 4$ and in $\hat{p}''$ it is the case that $i = 5$. According to our intuitions, $\hat{p}'$ and $\hat{p}''$ should be equally penalized. While NMD indeed penalizes them equally (since $\mathrm{NMD}(p, \hat{p}') = \mathrm{NMD}(p, \hat{p}'') = 0.1$), RNOD does not (since $\mathrm{RNOD}(p, \hat{p}') \approx 0.077$ while $\mathrm{RNOD}(p, \hat{p}'') \approx 0.092$).

[11] Sakai, T.: Comparing two binned probability distributions for information access evaluation. In: SIGIR. pp. 1073–1076 (2018)

Table 3. Average performance in terms of RNOD (lower is better), in analogy to the NMD results from Tab. 2. For each data set (Amazon-OQ-BK and Fact-OQ), we present the results of the two protocols APP and APP-OQ. The best performance in each column is highlighted in boldface. We further highlight all methods which are not statistically significantly different from the best method, as according to a Wilcoxon signed rank test with $p = 0.01$.

method	Amazon-OQ-BK		Fact-OQ	
	APP	APP-OQ	APP	APP-OQ
CC	.1151 ± .048	.0606 ± .020	.1319 ± .036	.1071 ± .027
PCC	.1360 ± .054	.0758 ± .025	.1372 ± .034	.1096 ± .026
ACC	.0487 ± .024	.0374 ± .016	.1563 ± .040	.1375 ± .030
PACC	.0419 ± .019	.0327 ± .014	.1750 ± .056	.1719 ± .047
SLD	**.0363 ± .017**	.0302 ± .014	.0890 ± .029	.0767 ± .021
OQT	.1542 ± .064	.0960 ± .032	.1456 ± .035	.1225 ± .032
ARC	.1303 ± .056	.0770 ± .027	.1242 ± .032	.0973 ± .022
IBU	.0534 ± .025	.0357 ± .014	**.0822 ± .028**	.0649 ± .018
RUN	.0531 ± .025	.0361 ± .014	.0869 ± .029	.0685 ± .019
o-ACC	.0487 ± .024	.0353 ± .014	.1032 ± .033	.0754 ± .016
o-PACC	.0419 ± .019	.0316 ± .012	.0914 ± .029	**.0625 ± .016**
o-SLD	**.0365 ± .017**	**.0296 ± .013**	.0857 ± .027	.0658 ± .015

Other OQ evaluation measures are proposed by Sakai[12], such as *Root Symmetric Normalized Order-aware Divergence* (RSNOD) and *Root Normalized Average Distance-Weighted sum of squares* (RNADW), but we do not consider them here since they are variants of RNOD that suffer from the same problem.

B.2 Results on other data sets

We have repeated our experiment from Tab. 2 with several other data sets.

First, we employ a different representation of the Amazon-OQ-BK data, namely a TFIDF representation, instead of the RoBERTa embeddings we employ in the main paper. To this end, we extract all uni- and bi-grams that appear at least 5 times in the training set and we use the logarithmic variant of the term-frequency factor, i.e., we compute the term-frequency as $1 + \log(tf)$. The aim of this experiment is to better understand how the quality of the representations (RoBERTa representations are assumed to be much more meaningful than TFIDF) affect the results. The results with this representation are presented in Tab. 4. They show that RoBERTa yields much better representations. However, having a wider margin for improvements in the TFIDF representation allows the ordered variants (o-ACC, o-PACC, o-SLD) to exhibit a more pronounced improvement with respect to the non-ordered variants (ACC, PACC, SLD); this finding is especially evident in the case of o-PACC.

Second, we evaluate on a collection of 4 public data sets from the UCI repository and OpenML. To this end, we have first selected regression data sets with at least 30 000 items. From there on, we have tried to find an equidistant binning

[12] Sakai, T.: A closer look at evaluation measures for ordinal quantification. In: CIKM 2021 Workshop on Learning to Quantify (2021)

which produces at least 10 bins (i.e., ordered classes), each of which having at least 1000 items. We only retain data sets for which such a binning was possible, and we remove all items that lie outside the 10 equidistant bins. In order to retain as many samples as possible, we maximize the distance between the left-most and right-most bin boundaries. If less than 30 000 items remain, we discard the data set. From this protocol, we obtain the 4 data sets Uci-blog-feedback-OQ, Uci-online-news-popularity-OQ, OpenMl-Yolanda-OQ, and OpenMl-fried-OQ. We present the results obtained with these data sets in Tab. 5. They confirm our main conclusions: the ordered modifications consistently improve over the original (non-ordered) versions, with o-PACC and IBU often displaying the best overall results.

Table 4. NMD on a TFIDF representation, instead of RoBERTa embeddings, of the Amazon-OQ-BK data set.

method	Amazon-OQ-BK (TFIDF)	
	APP	APP-OQ
CC	.0867 ± .034	.0683 ± .031
PCC	.1082 ± .044	.0950 ± .048
ACC	.0353 ± .015	.0333 ± .014
PACC	.0301 ± .015	.0310 ± .015
SLD	.0477 ± .018	.0381 ± .012
OQT	.1583 ± .065	.1539 ± .072
ARC	.0989 ± .037	.0855 ± .038
IBU	.0596 ± .023	.0454 ± .020
RUN	.0594 ± .023	.0452 ± .020
o-ACC	.0347 ± .017	.0227 ± .009
o-PACC	**.0276 ± .014**	**.0194 ± .007**
o-SLD	.0477 ± .018	.0363 ± .011

Table 5. NMD in additional datasets

method	Uci-blog-feedback-OQ		Uci-online-news-popularity-OQ		OpenMl-Yolanda-OQ		OpenMl-fried-OQ	
	APP	APP-OQ	APP	APP-OQ	APP	APP-OQ	APP	APP-OQ
CC	.0958 ± .034	.0884 ± .031	.1664 ± .047	.1549 ± .045	.0767 ± .023	.0779 ± .025	.0330 ± .008	.0243 ± .006
PCC	.0967 ± .042	.0960 ± .045	.0996 ± .044	.0985 ± .047	.0926 ± .030	.0921 ± .032	.0410 ± .010	.0330 ± .008
ACC	.1147 ± .042	.1144 ± .045	.1365 ± .055	.1357 ± .060	.0807 ± .024	.0824 ± .026	.0454 ± .021	.0482 ± .023
PACC	.1323 ± .049	.1437 ± .050	.1515 ± .063	.1246 ± .055	.1068 ± .047	.1102 ± .050	.0614 ± .026	.0659 ± .026
SLD	.1001 ± .044	.1224 ± .038	.1576 ± .063	.1687 ± .069	.0753 ± .025	.0784 ± .028	.0369 ± .009	.0373 ± .008
OQT	.2222 ± .058	.2050 ± .057	.3220 ± .087	.3177 ± .092	.2246 ± .056	.2223 ± .058	.0566 ± .014	.0472 ± .012
ARC	.2420 ± .062	.2474 ± .063	.3801 ± .085	.3793 ± .089	.2513 ± .058	.2500 ± .060	.0589 ± .017	.0598 ± .018
IBU	.0997 ± .046	.0980 ± .049	.0886 ± .039	.0858 ± .043	**.0558 ± .017**	.0553 ± .018	**.0168 ± .005**	**.0146 ± .004**
RUN	.1348 ± .052	.1339 ± .054	.1115 ± .048	.1181 ± .053	.0577 ± .017	.0604 ± .018	.0206 ± .006	.0161 ± .005
o-ACC	.0772 ± .031	.0728 ± .027	**.0833 ± .030**	**.0718 ± .027**	.0568 ± .016	.0549 ± .017	.0264 ± .008	.0189 ± .004
o-PACC	**.0747 ± .028**	**.0664 ± .025**	.0954 ± .039	.0804 ± .031	.0580 ± .014	**.0537 ± .014**	.0350 ± .018	**.0146 ± .004**
o-SLD	.1195 ± .041	.1190 ± .040	.0993 ± .044	.0992 ± .046	.0701 ± .019	.0648 ± .019	.0322 ± .007	.0282 ± .005

B.3 Hyperparameter grids

In our experiments, each method has the opportunity to optimize its hyperparameters on the APP and on the APP-OQ validation samples. These hyperparameters consist of parameters of the quantifier and of parameters of the classifier, with which the quantifier is equipped. After taking out preliminary experiments, which we omit here for conciseness, we have chosen different hyperparameter grids for the different data sets.

To this end, Tab. 6 and Tab. 7 present the parameters for the Amazon-OQ-BK data set. For instance, CC and PCC can choose between 10 hyperparameter configurations of the classifier (2 class weights × 5 regularization parameters), but they do not have additional parameters on the quantification level. We note that an inspection of the validation results revealed that the fraction of held-out data does not considerably affect the results of ACC, PACC, OQT, and ARC. Therefore, for OQT and ARC we decided to fix the proportion of the held-out split to 1/3 and do not include this hyperparameter in the exploration, since those methods are computationally expensive.

Tab. 8 and Tab. 9 present the parameters for the Fact-OQ data. For conciseness, they also contain the parameters for the UCI and OpenML data sets. The remaining parameters for the UCI and OpenML data sets are presented in Tab. 10.

Table 6. Hyperparameter grid of classifiers when analyzing the Amazon-OQ-BK data in the experiment from Tab. 2.

classifier	parameter	values
logistic regression	class weight	{balanced, unbalanced }
	regularization parameter C	$\{0.001, 0.01, 0.1, 1.0, 10.0\}$

B.4 Performance in other APP plausibility levels

Our APP-OQ protocol selects the 20% of validation and test samples which we deem most plausible. For completeness, we include here the results for other plausibility levels, which are the second-most, the third-most, the fourth-most, and the least plausible 20%. In other words: we have divided all APP samples in terms of their conceived plausibility into five levels, the first of which makes our APP-OQ, and we have evaluated all methods in all of these plausibility levels. Recall that every evaluation entails the optimization of the hyperparameters on the corresponding validation split. Recall also that every split concerns only the validation and the test data, while the training set is the same in all cases.

In order to make our experimentation more transparent, we report the results accompanied by the hyperparameters that each method has chosen on the validation samples. The following tables only consider NMD, but the LaTeX sources of the RNOD tables are made available as part of our supplementary material as well. Often, the regularization hyperparameters favouring smoother solutions are stronger in the smoothest cases.

Table 7. Hyperparameter grid of quantification methods when analyzing the Amazon-OQ-BK data in the experiment from Tab. 2.

method	parameter	values
CC	no parameters	
PCC	no parameters	
ACC	fraction of held-out data	$\{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}\}$
PACC	fraction of held-out data	$\{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}\}$
SLD	no parameters	
OQT	fraction of held-out data	$\{\frac{1}{3}\}$
ARC	fraction of held-out data	$\{\frac{1}{3}\}$
RUN	τ	{3e-2, 1e-2, 3e-3, 1e-3, 3e-4, 1e-4, 3e-5, 1e-6}
IBU	order of polynomial	$\{0, 1, 2\}$
	interpolation factor	{3e-1, 1e-1, 3e-2, 1e-2, 3e-3, 1e-3}
o-ACC	fraction of held-out data	$\{\frac{1}{4}, \frac{1}{3}\}$
	τ	{1e-2, 3e-3, 1e-3, 3e-4, 1e-4, 1e-5, 1e-6, 1e-9}
o-PACC	fraction of held-out data	$\{\frac{1}{4}, \frac{1}{3}\}$
	τ	{1e-2, 3e-3, 1e-3, 3e-4, 1e-4, 1e-5, 1e-6, 1e-9}
o-SLD	order of polynomial	$\{0, 1, 2\}$
	interpolation factor	{1e-1, 3e-2, 1e-2, 3e-3, 1e-3}

Table 8. Hyperparameter grid of classifiers when analyzing the Fact-OQ data in the experiment from Tab. 2.

classifier	parameter	values
probability-calibrated decision tree	class weight	{balanced, unbalanced}
	split criterion	{Gini index, Entropy}
	maximum depth	$\{4, 6, 8, 10, 12\}$

Table 9. Hyperparameter grid of quantification methods when analyzing the Fact-OQ data in the experiment from Tab. 2 or any of the data sets Uci-blog-feedback-OQ, Uci-online-news-popularity-OQ, OpenMl-Yolanda-OQ, and OpenMl-fried-OQ.

method	parameter	values
CC	no parameters	
PCC	no parameters	
ACC	fraction of held-out data	$\{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}\}$
PACC	fraction of held-out data	$\{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}\}$
SLD	no parameters	
OQT	fraction of held-out data	$\{\frac{1}{3}\}$
ARC	fraction of held-out data	$\{\frac{1}{3}\}$
RUN	τ	{1e-1, 1e-3, 1e-5}
	number of leaf nodes	$\{60, 120, 180\}$
IBU	order of polynomial	$\{0, 1, 2\}$
	interpolation factor	{0.1, 0.01, 0.0}
	number of leaf nodes	$\{60, 120, 180\}$
o-ACC	fraction of held-out data	$\{\frac{1}{3}\}$
	τ	{1e-1, 1e-3, 1e-5}
o-PACC	fraction of held-out data	$\{\frac{1}{3}\}$
	τ	{1e-1, 1e-3, 1e-5}
o-SLD	order of polynomial	$\{0, 1, 2\}$
	interpolation factor	{1e-1, 3e-2, 1e-2}

Table 10. Hyperparameter grid of classifiers when analyzing any of the data sets Uci-blog-feedback-OQ, Uci-online-news-popularity-OQ, OpenMl-Yolanda-OQ, and OpenMl-fried-OQ.

classifier	parameter	values
probability-calibrated decision tree	class weight	{balanced, unbalanced}
	split criterion	{Gini index, Entropy}
	maximum depth	$\{4, 6, 8, 10, 12\}$
logistic regression	class weight	{balanced, unbalanced}
	regularization parameter C	$\{0.001, 0.01, 0.1, 1.0, 10.0\}$

Table 11. NMD on Amazon-OQ-BK, level 2 out of 5

quantification method	avg. NMD ± stddev.
SLD on LR ($w = n, C = 0.01$)	$\mathbf{0.0164 \pm 0.0061}$
o-SLD ($o = 2, i = 0.001$) on LR ($w = n, C = 0.01$)	0.0164 ± 0.0061
PACC ($v = \frac{1}{4}$) on LR ($w = u, C = 0.1$)	0.0190 ± 0.0070
o-PACC ($r = I, \tau = 1.0e-9, v = \frac{1}{4}$) on LR ($w = u, C = 0.1$)	0.0190 ± 0.0070
ACC ($v = \frac{1}{4}$) on LR ($w = u, C = 0.1$)	0.0210 ± 0.0077
o-ACC ($r = I, \tau = 1.0e-9, v = \frac{1}{4}$) on LR ($w = u, C = 0.1$)	0.0210 ± 0.0077
RUN ($\tau = 1.0e-6$) on LR ($w = u, C = 1.0$)	0.0221 ± 0.0079
IBU ($o = 2, i = 0.001$) on LR ($w = u, C = 1.0$)	0.0222 ± 0.0079
CC on LR ($w = u, C = 10.0$)	0.0423 ± 0.0122
PCC on LR ($w = u, C = 10.0$)	0.0524 ± 0.0156
ARC ($v = \frac{1}{3}$) on LR ($w = u, C = 1.0$)	0.0527 ± 0.0168
OQT ($v = \frac{1}{3}$) on LR ($w = n, C = 10.0$)	0.0654 ± 0.0225

Table 12. NMD on Amazon-OQ-BK, level 3 out of 5

quantification method	avg. NMD ± stddev.
SLD on LR ($w = n, C = 0.01$)	$\mathbf{0.0172 \pm 0.0066}$
o-SLD ($o = 0, i = 0.01$) on LR ($w = n, C = 0.001$)	$\mathbf{0.0174 \pm 0.0076}$
PACC ($v = \frac{1}{4}$) on LR ($w = u, C = 0.1$)	0.0199 ± 0.0077
o-PACC ($r = I, \tau = 1.0e-9, v = \frac{1}{4}$) on LR ($w = u, C = 0.1$)	0.0199 ± 0.0077
ACC ($v = \frac{1}{3}$) on LR ($w = u, C = 10.0$)	0.0218 ± 0.0085
o-ACC ($r = I, \tau = 1.0e-9, v = \frac{1}{3}$) on LR ($w = u, C = 10.0$)	0.0218 ± 0.0085
RUN ($\tau = 1.0e-6$) on LR ($w = u, C = 0.001$)	0.0244 ± 0.0089
IBU ($o = 2, i = 0.001$) on LR ($w = u, C = 0.001$)	0.0246 ± 0.0089
CC on LR ($w = u, C = 10.0$)	0.0503 ± 0.0116
PCC on LR ($w = u, C = 10.0$)	0.0603 ± 0.0146
ARC ($v = \frac{1}{3}$) on LR ($w = u, C = 1.0$)	0.0604 ± 0.0179
OQT ($v = \frac{1}{3}$) on LR ($w = n, C = 1.0$)	0.0738 ± 0.0231

Table 13. NMD on Amazon-OQ-BK, level 4 out of 5

quantification method	avg. NMD ± stddev.
o-SLD ($o = 0, i = 0.01$) on LR ($w = n, C = 0.001$)	$\mathbf{0.0177 \pm 0.0072}$
SLD on LR ($w = n, C = 0.01$)	0.0178 ± 0.0068
PACC ($v = \frac{1}{3}$) on LR ($w = u, C = 0.01$)	0.0215 ± 0.0081
o-PACC ($r = I, \tau = 1.0e-9, v = \frac{1}{3}$) on LR ($w = u, C = 0.01$)	0.0215 ± 0.0081
ACC ($v = \frac{1}{3}$) on LR ($w = u, C = 10.0$)	0.0238 ± 0.0093
o-ACC ($r = I, \tau = 1.0e-9, v = \frac{1}{3}$) on LR ($w = u, C = 10.0$)	0.0238 ± 0.0093
RUN ($\tau = 1.0e-6$) on LR ($w = u, C = 0.01$)	0.0267 ± 0.0091
IBU ($o = 2, i = 0.001$) on LR ($w = u, C = 0.01$)	0.0269 ± 0.0091
CC on LR ($w = u, C = 1.0$)	0.0595 ± 0.0116
ARC ($v = \frac{1}{3}$) on LR ($w = u, C = 1.0$)	0.0695 ± 0.0172
PCC on LR ($w = u, C = 10.0$)	0.0700 ± 0.0139
OQT ($v = \frac{1}{3}$) on LR ($w = n, C = 1.0$)	0.0823 ± 0.0219

Table 14. NMD on Amazon-OQ-BK, level 5 out of 5 (the least smooth)

quantification method	avg. NMD ± stddev.
o-SLD ($o = 0, i = 0.01$) on LR ($w = n, C = 0.001$)	$\mathbf{0.0177 \pm 0.0071}$
SLD on LR ($w = n, C = 0.01$)	0.0193 ± 0.0073
PACC ($v = \frac{1}{4}$) on LR ($w = n, C = 0.1$)	0.0234 ± 0.0081
o-PACC ($r = I, \tau = 1.0e - 9, v = \frac{1}{4}$) on LR ($w = n, C = 0.1$)	0.0234 ± 0.0081
ACC ($v = \frac{1}{3}$) on LR ($w = u, C = 10.0$)	0.0286 ± 0.0106
o-ACC ($r = I, \tau = 1.0e - 9, v = \frac{1}{3}$) on LR ($w = u, C = 10.0$)	0.0286 ± 0.0106
RUN ($\tau = 1.0e - 6$) on LR ($w = u, C = 0.001$)	0.0328 ± 0.0105
IBU ($o = 0, i = 0.001$) on LR ($w = u, C = 0.001$)	0.0329 ± 0.0105
CC on LR ($w = u, C = 1.0$)	0.0761 ± 0.0135
PCC on LR ($w = u, C = 10.0$)	0.0878 ± 0.0158
ARC ($v = \frac{1}{3}$) on LR ($w = u, C = 0.1$)	0.0895 ± 0.0166
OQT ($v = \frac{1}{3}$) on LR ($w = n, C = 0.01$)	0.1023 ± 0.0193

Table 15. NMD on Fact-OQ, level 2 out of 5

quantification method	avg. NMD ± stddev.
IBU ($o = 0, i = 0.01, J = 60$)	$\mathbf{0.0199 \pm 0.0047}$
o-PACC ($r = I, \tau = 0.001, v = \frac{1}{3}$) on DT ($w = n, c = E, d = 8$)	0.0203 ± 0.0039
RUN ($\tau = 1.0e - 5, J = 60$)	0.0205 ± 0.0049
o-ACC ($r = C_2, \tau = 0.001, v = \frac{1}{3}$) on DT ($w = u, c = G, d = 8$)	0.0248 ± 0.0060
o-SLD ($o = 0, i = 0.03$) on DT ($w = n, c = E, d = 4$)	0.0307 ± 0.0068
SLD on DT ($w = n, c = G, d = 6$)	0.0359 ± 0.0091
CC on DT ($w = u, c = G, d = 8$)	0.0506 ± 0.0112
ARC ($v = \frac{1}{3}$) on DT ($w = u, c = G, d = 8$)	0.0556 ± 0.0147
ACC ($v = \frac{1}{4}$) on DT ($w = n, c = G, d = 10$)	0.0585 ± 0.0285
PCC on DT ($w = u, c = E, d = 6$)	0.0623 ± 0.0170
OQT ($v = \frac{1}{3}$) on DT ($w = u, c = G, d = 6$)	0.0728 ± 0.0197
PACC ($v = \frac{1}{4}$) on DT ($w = n, c = G, d = 4$)	0.0802 ± 0.0298

Table 16. NMD on Fact-OQ, level 3 out of 5

quantification method	avg. NMD ± stddev.
IBU ($o = 2, i = 0.01, J = 60$)	$\mathbf{0.0210 \pm 0.0049}$
RUN ($\tau = 1.0e - 5, J = 60$)	0.0217 ± 0.0050
o-PACC ($r = I, \tau = 0.001, v = \frac{1}{3}$) on DT ($w = n, c = E, d = 8$)	0.0225 ± 0.0039
o-ACC ($r = C_2, \tau = 0.001, v = \frac{1}{3}$) on DT ($w = u, c = G, d = 8$)	0.0267 ± 0.0060
o-SLD ($o = 0, i = 0.03$) on DT ($w = n, c = E, d = 4$)	0.0326 ± 0.0068
SLD on DT ($w = n, c = G, d = 6$)	0.0374 ± 0.0095
CC on DT ($w = u, c = G, d = 8$)	0.0523 ± 0.0105
ARC ($v = \frac{1}{3}$) on DT ($w = u, c = G, d = 8$)	0.0562 ± 0.0141
ACC ($v = \frac{1}{4}$) on DT ($w = n, c = G, d = 10$)	0.0579 ± 0.0285
PCC on DT ($w = u, c = E, d = 6$)	0.0644 ± 0.0160
OQT ($v = \frac{1}{3}$) on DT ($w = u, c = G, d = 6$)	0.0744 ± 0.0193
PACC ($v = \frac{1}{3}$) on DT ($w = n, c = G, d = 10$)	0.0785 ± 0.0481

Table 17. NMD on Fact-OQ, level 4 out of 5

quantification method	avg. NMD ± stddev.
IBU ($o = 0, i = 0.01, J = 60$)	$\mathbf{0.0224 \pm 0.0052}$
RUN ($\tau = 1.0e - 5, J = 60$)	0.0234 ± 0.0052
o-PACC ($r = I, \tau = 0.001, v = \frac{1}{3}$) on DT ($w = n, c = E, d = 8$)	0.0251 ± 0.0040
o-ACC ($r = C_2, \tau = 0.001, v = \frac{1}{3}$) on DT ($w = u, c = G, d = 8$)	0.0292 ± 0.0064
o-SLD ($o = 0, i = 0.03$) on DT ($w = n, c = E, d = 4$)	0.0342 ± 0.0069
SLD on DT ($w = n, c = G, d = 6$)	0.0380 ± 0.0094
CC on DT ($w = u, c = G, d = 8$)	0.0543 ± 0.0110
ARC ($v = \frac{1}{3}$) on DT ($w = u, c = G, d = 8$)	0.0561 ± 0.0138
ACC ($v = \frac{1}{4}$) on DT ($w = n, c = G, d = 10$)	0.0582 ± 0.0277
PCC on DT ($w = u, c = E, d = 6$)	0.0653 ± 0.0162
OQT ($v = \frac{1}{3}$) on DT ($w = u, c = G, d = 6$)	0.0745 ± 0.0184
PACC ($v = \frac{1}{4}$) on DT ($w = u, c = E, d = 12$)	0.0788 ± 0.0320

Table 18. NMD on Fact-OQ, level 5 out of 5 (the least smooth)

quantification method	avg. NMD ± stddev.
IBU ($o = 1, i = 0.0, J = 60$)	$\mathbf{0.0245 \pm 0.0067}$
RUN ($\tau = 1.0e - 5, J = 60$)	0.0262 ± 0.0058
o-PACC ($r = I, \tau = 0.001, v = \frac{1}{3}$) on DT ($w = u, c = G, d = 10$)	0.0298 ± 0.0049
o-ACC ($r = C_2, \tau = 0.001, v = \frac{1}{3}$) on DT ($w = u, c = G, d = 10$)	0.0330 ± 0.0062
o-SLD ($o = 0, i = 0.01$) on DT ($w = n, c = E, d = 4$)	0.0368 ± 0.0096
SLD on DT ($w = n, c = E, d = 6$)	0.0393 ± 0.0112
ARC ($v = \frac{1}{3}$) on DT ($w = u, c = G, d = 8$)	0.0583 ± 0.0131
CC on DT ($w = u, c = G, d = 8$)	0.0604 ± 0.0129
ACC ($v = \frac{1}{3}$) on DT ($w = n, c = E, d = 8$)	0.0646 ± 0.0274
PCC on DT ($w = u, c = E, d = 6$)	0.0715 ± 0.0188
PACC ($v = \frac{1}{3}$) on DT ($w = n, c = G, d = 10$)	0.0776 ± 0.0455
OQT ($v = \frac{1}{3}$) on DT ($w = u, c = G, d = 6$)	0.0783 ± 0.0193