

Reducing High-Dimensional Data by Principal Component Analysis vs. Random Projection for Nearest Neighbor Classification

Sampath Deegalla*

*Dept. of Computer and Systems Sciences,
Stockholm University and Royal Institute of Technology,
Forum 100, SE-164 40 Kista, Sweden.
si-sap@dsv.su.se

Henrik Boström†

†Dept. of Computer and Systems Sciences,
Stockholm University and Royal Institute of Technology,
Forum 100, SE-164 40 Kista, Sweden.
henke@dsv.su.se

Abstract

The computational cost of using nearest neighbor classification often prevents the method from being applied in practice when dealing with high-dimensional data, such as images and micro arrays. One possible solution to this problem is to reduce the dimensionality of the data, ideally without losing predictive performance. Two different dimensionality reduction methods, principal component analysis (PCA) and random projection (RP), are compared w.r.t. the performance of the resulting nearest neighbor classifier on five image data sets and two micro array data sets. The experimental results show that PCA results in higher accuracy than RP for all the data sets used in this study. However, it is also observed that RP generally outperforms PCA for higher numbers of dimensions. This leads to the conclusion that PCA is more suitable in time-critical cases (i.e., when distance calculations involving only a few dimensions can be afforded), while RP can be more suitable when less severe dimensionality reduction is required. In 6 respectively 4 cases out of 7, the use of PCA and RP even outperform using the non-reduced feature set, hence not only resulting in more efficient, but also more effective, nearest neighbor classification.

1 Introduction

With the development of technology, large volumes of high-dimensional data become rapidly available and easily accessible for the data mining community. Such data includes high resolution images, text documents, and gene expressions data and so on. However, high dimensional data puts demands on the learning algorithm both in terms of efficiency and effectiveness. The *curse of dimensionality* is a well known phenomenon that occurs when the generation of a predictive model is misled by an overwhelming number of features to choose between, e.g., when deciding what feature to use in a node of a decision tree (Witten and Frank, 2005). Some learning methods are less sensitive to this problem since they do not rely on choosing a subset of the features, but instead base the classification on all available features. Nearest

neighbor classifiers belong to this category of methods (Witten and Frank, 2005). However, although increasing the number of dimensions does not typically have a detrimental effect on predictive performance, the computational cost may be prohibitively large, effectively preventing the method from being used in many cases with high-dimensional data.

In this work, we consider two methods for dimensionality reduction, principal component analysis (PCA) and random projection (RP) (Bingham and Mannila, 2001; Fradkin and Madigan, 2003; Fern and Brodley, 2003; Kaski, 1998). We investigate which of these is most suited for being used in conjunction with nearest neighbor classification when dealing with two types of high-dimensional data: images and micro arrays.

In the next section, we provide a brief description of PCA and RP and compare them w.r.t. computa-

tional complexity. In section three, we present results from a comparison of the methods when used together with nearest neighbor classification on five image data sets and two micro array data sets, together with an analysis of the results. In section four, we discuss some related work, and finally, in section five, we give some concluding remarks and point out some directions for future work.

2 Dimensionality reduction methods

Principal component analysis (PCA) and Random projection (RP) are two dimensionality reduction methods that have been used successfully in conjunction with learning methods (Bingham and Mannila, 2001; Fradkin and Madigan, 2003). PCA is the most well-known and popular of the above two, whereas RP is more recently gaining popularity among researchers (Bingham and Mannila, 2001; Fradkin and Madigan, 2003; Fern and Brodley, 2003; Kaski, 1998), not least by being much more efficient.

Principal component analysis (PCA)

PCA is a technique which uses a linear transformation to form a simplified data set retaining the characteristics of the original data set.

Assume that the original data matrix contains d dimensions and n observations and it is required to reduce the dimensionality into a k dimensional subspace. This transformation is given by

$$Y = E^T X \quad (1)$$

Here $E_{d \times k}$ is the projection matrix which contains k eigen vectors corresponds to the k highest eigen values, and where $X_{d \times n}$ is a mean centered data matrix. We have followed the Singular Value Decomposition (SVD) of the data matrix to calculate principal components.

Random projection(RP)

Random projection is based on matrix manipulation, which uses a random matrix to project the original data set into a low dimensional subspace (Bingham and Mannila, 2001; Fradkin and Madigan, 2003).

Assume that it is required to reduce the d dimensional data set into a k dimensional set where the number of instances are n ,

$$Y = R X \quad (2)$$

Here $R_{k \times d}$ is the random matrix and $X_{d \times n}$ is the original matrix. The idea for random projection originates from the Johnson-Lindenstrauss lemma (JL) (Dasgupta and Gupta, 1999). It states that a set of n points could be projected from $R^d \rightarrow R^k$ while approximately preserving the Euclidean distance between the points within an arbitrarily small factor. For the theoretical effectiveness of random projection method, see (Fradkin and Madigan, 2003).

Several algorithms have been proposed to generate the matrix R according to the JL, and Achlioptas's way of constructing R have received much attention in the literature (Bingham and Mannila, 2001; Fradkin and Madigan, 2003). According to Achlioptas (Achlioptas, 2001), the elements of the random matrix R can be constructed in the following way:

$$r_{ij} = \begin{cases} +\sqrt{3} & \text{with } P_r = \frac{1}{6}; \\ 0 & \text{with } P_r = \frac{2}{3}; \\ -\sqrt{3} & \text{with } P_r = \frac{1}{6}. \end{cases} \quad (3)$$

An analysis of the computational complexity of random projection shows that it is very efficient compared to principal component analysis. Random projection requires only $O(dkn)$ whereas principal component analysis needs $O(d^2n) + O(d^3)$ (Bingham and Mannila, 2001). If rank of matrix X is r then computational complexity for SVD is $O(drn)$ (Bingham and Mannila, 2001).

3 Empirical study

3.1 Experimental setup

In comparison of the use of PCA and RP for nearest neighbor classification, five image data sets (IRMA (Lehmann et al., 2000), COIL-100 (Nene et al., 1996), ZuBuD (H. Shao et al., 2003), MIAS (MIAS data set), Outex (Ojala et al., 2002)) and two micro array data sets (Colon Tumor (Alon et al., 1999), Leukemia (Golub et al., 1999)) have been used. The image data sets consist of two medical image data sets (IRMA, MIAS), two object recognition data sets (COIL-100, ZuBuD) and a texture analysis data set (Outex - *TC_00013*). The IRMA (Image Retrieval and Medical Application) data set contains radiography images of 57 classes, where the quality of the images varies significantly. The COIL-100 (Columbia university image library) data set consists of images of 100 objects, while ZuBuD (Zurich Building Image Database) contains images of 201 buildings in Zurich city. MIAS (The Mammography Image Analysis Society) mini mammography database contains mam-

Table 1: Description of data

Data set	Instances	Attributes	# of Classes
IRMA	9000	1024	57
COIL100	7200	1024	100
ZuBuD	1005	1024	201
MIAS	322	1024	7
Outex	680	1024	68
Colon Tumor	62	2000	2
Leukemia	38	7129	2

mography images of 7 categories and finally Outex (University of Oulu Texture Database) image data set contains images of 68 general textures. Two micro array data sets are also included in the study: Leukemia (ALL-AML¹) (Golub et al., 1999) data set and Colon Tumor (Alon et al., 1999).

For all image data sets, colour images have been converted into gray scale images with 256 gray values and then resized into 32×32 pixel sized images. The original matrices consist of pixel brightness values as features. Therefore, all image data sets contain 1024 attributes. The number of attributes for the micro array data sets and the number of instances for all data sets are shown in Table 1.

The Waikato Environment for Knowledge Analysis (WEKA) (Witten and Frank, 2005) implementation of the nearest neighbor classifier was used for this study. PCA was done using MATLAB whereas WEKA's RP filter implementation was used for RP. The accuracies were estimated using ten-fold cross-validation, and the results for RP is the average from 30 runs to account for its random nature.

3.2 Experimental results

The accuracies of using a nearest neighbor classifier on data reduced by PCA and RP, as well as without dimensionality reduction, are shown in Figure 1 for various number of resulting dimensions. Table 2 shows the highest accuracies obtained by using the two dimensionality reduction methods with the corresponding number of dimensions. Table 3 shows the accuracies obtained for PCA and Table 4 shows the accuracies and standard deviations (SD) for RP.

¹ALL: acute lymphoblastic leukemia
AML: acute myeloid leukemia

3.3 Analysis

The experimental results show that reducing the dimensionality by using PCA results in a higher accuracy than by using RP for all data sets used in this study. In Table 2, it can be seen that only a few principal components is required for achieving the highest accuracy. However, RP typically requires a larger number of dimensions compared to PCA to obtain a high accuracy.

Classification accuracy using PCA typically has its peak for a small number of dimensions, after which the accuracy degrades. In contrast to this, the accuracy of RP generally increases with the number of dimensions. For large numbers of dimensions, RP generally outperforms PCA.

It can also be seen that the time required for doing a prediction is reduced when using a dimensionality reduction method as shown in Table 5. Even though nearest neighbor classification is often less efficient when compared to other learning methods, the time required may be significantly reduced by using the dimensionality reduction while still maintaining a high accuracy.

The results can be interpreted as if one can only afford a few number of dimensions (e.g. due to time constraints during classification), PCA appears to be more suitable than RP. On the other hand, for large numbers of dimensions, RP appears more suitable.

It can be noted that in 6 respectively 4 cases out of 7, the use of PCA and RP, even outperform using the non-reduced feature set, hence demonstrating that dimensionality reduction may not only lead to more efficient, but also more effective, nearest neighbor classification.

4 Related work

Fradkin and Madigan (Fradkin and Madigan, 2003) have compared PCA and RP with decision trees (C4.5), k-nearest-neighbor classification with $k=1$ and $k=5$ and support vector machines for supervised learning. In their study, PCA outperforms RP, but it was also realized that there was a significant computational overhead of using PCA compared using RP.

Bingham and Mannila (Bingham and Mannila, 2001) have also compared RP with several other dimensionality reduction methods such as PCA, singular value decomposition (SVD), Latent semantic indexing (LSI) and Discrete cosine transform (DCT) for image and text data. The criteria chosen for the comparison was the amount of distortion caused on the original data and the computational complexity.

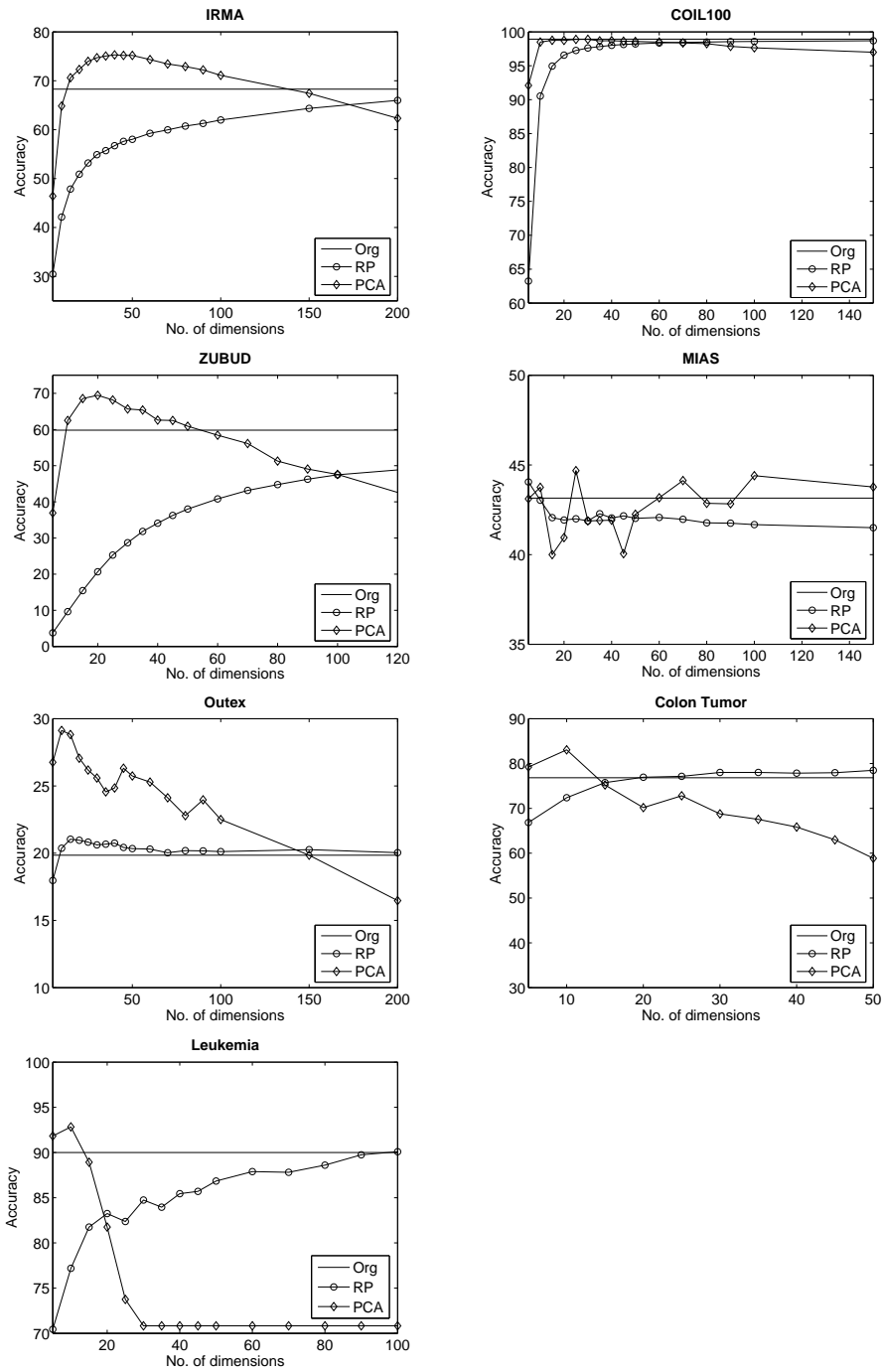


Figure 1: Comparison of the accuracies among Original and PCA and RP based attributes.

Table 2: The highest prediction accuracy (%) of PCA and RP with minimum number of dimensions

	IRMA		COIL-100		ZuBuD		MIAS		Outex		Colon		Leukemia	
PCA	75.30	(40)	98.90	(30)	69.46	(20)	53.76	(250)	29.12	(10)	83.05	(10)	92.83	(10)
RP	67.01	(250)	98.79	(250)	54.01	(250)	44.05	(5)	21.04	(15)	80.22	(150,200)	91.32	(150)
All	68.29	(1024)	98.92	(1024)	59.81	(1024)	43.15	(1024)	19.85	(1024)	76.83	(2000)	90.00	(7129)

They also extended their experiments to determine the effects on noisy images and noiseless images.

Fern and Brodley (Fern and Brodley, 2003) have used Random projections for unsupervised learning. They have experimented with using RP for clustering of high dimensional data using multiple random projections with ensemble methods. Furthermore, they also proposed a novel approach based on RP for clustering, which was compared with PCA.

Kaski (Kaski, 1998) used RP in the WEBSOM system for document clustering. RP was compared to PCA for reducing the dimensionality of the data in order to construct Self-Organized Maps.

5 Concluding remarks

We have compared the use of PCA and RP for reducing dimensionality of data to be used by a nearest neighbor classifier on five image data sets and two micro array data sets. It was observed that for all data sets, the use of PCA resulted in a higher accuracy compared to using RP. It was also observed that PCA is more effective for severe dimensionality reduction, while RP is more suitable when keeping a high number of dimensions (although a high number is not always optimal w.r.t. accuracy).

In 6 cases out of 7 for PCA, and in 4 cases out of 7 for RP, a higher accuracy was obtained than using nearest-neighbor classification with the non-reduced feature set. This shows that dimensionality reduction may not only lead to more efficient, but also more effective, nearest neighbor classification.

Directions for future work include considering other types of high-dimensional data to gain a further understanding of the type of data for which each of the two dimensionality reduction techniques is best suited, as well as considering other dimensionality reduction methods for nearest neighbor classification.

Acknowledgements

The authors would like to thank T.M. Lehmann, Dept. of Medical Informatics, RWTH Aachen, Germany for providing the "10000 IRMA images of 57 categories".

Financial support for the first author by SIDA/SAREC is greatly acknowledged.

References

- D. Achlioptas. Database-friendly random projections. In *ACM Symposium on the Principles of Database Systems*, pages 274–281, 2001.
- U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. In *Proc. Natl. Acad. Sci. USA*, volume 96, pages 6745–6750, 1999. Data set : <http://sdmc.lit.org.sg/GEDatasets/Datasets.html>.
- E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Knowledge Discovery and Data Mining*, pages 245–250, 2001.
- S. Dasgupta and A. Gupta. An elementary proof of the Johnson-Lindenstrauss lemma. Technical Report TR-99-006, International Computer Science Institute, Berkeley, CA, 1999.
- X. Z. Fern and C. E. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proceedings of the Twentieth International Conference of Machine Learning*, 2003.
- D. Fradkin and D. Madigan. Experiments with Random Projections for Machine Learning. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 517–522, 2003.

- T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999. Data set : <http://sdmc.lit.org.sg/GEDatasets/Datasets.html>.
- H. Shao, T. Svoboda, and L. Van Gool. Zubud - zurich building database for image based recognition. Technical report, Computer Vision Lab, Swiss Federal Institute of Technology, Switzerland, 2003. Data set : <http://www.vision.ee.ethz.ch/showroom/zubud/-index.en.html>.
- S. Kaski. Dimensionality Reduction by Random Mapping: Fast Similarity Computation for Clustering. In *Proceedings of IJCNN'98, International Joint Conference on Neural Networks*, volume 1, pages 413–418, Piscataway, NJ, 1998. IEEE Service Center.
- T. M. Lehmann, B. B. Wein, J. Dahmen, J. Bredno, F. Vogelsang, and M. Kohnen. Content-based image retrieval in medical applications: a novel multistep approach. In M. M. Yeung, B.-L. Yeo, and C. A. Bouman, editors, *Proceedings of SPIE: Storage and Retrieval for Media Databases 2000*, volume 3972, pages 312–320, 2000. Data set : <http://phobos.imib.rwth-aachen.de/irma/datasets.en.php>.
- MIAS data set. MIAS data set : <http://www.wiau.man.ac.uk/services/MIAS/-MIASmini.html>.
- S. A. Nene, S. K. Nayar, and H. Murase. Columbia Object Image Library: COIL-100. Technical report, CUCS-006-96, February 1996. Data set : <http://www1.cs.columbia.edu/CAVE/research/-softlib/coil-100.html>.
- T. Ojala, T. Maenpaa, M. Pietikainen, J. Viertola, J. Kyllonen, and S. Huovinen. Outex - new framework for empirical evaluation of texture analysis algorithms. In *ICPR '02: Proceedings of the 16th International Conference on Pattern Recognition (ICPR'02) Volume 1*, 2002. Data set : <http://www.outex.oulu.fi>.
- I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2005.

Table 3: Classification accuracies (%) on PCA

Dim. (k)	IRMA	COIL-100	ZuBuD	MIAS	Outex	Colon	Leukemia
5	46.41	92.10	36.92	43.11	26.76	79.21	91.83
10	64.83	98.50	62.50	43.75	29.12	83.05	92.83
15	70.62	98.75	68.56	40.00	28.82	75.21	88.92
20	72.31	98.74	69.46	40.95	27.06	70.17	81.75
25	73.97	98.88	68.16	44.68	26.18	72.81	73.75
30	74.71	98.90	65.68	41.88	25.59	68.76	70.83
35	75.06	98.68	65.38	41.90	24.56	67.55	70.83
40	75.30	98.71	62.60	41.92	24.85	65.83	70.83
45	75.18	98.63	62.50	40.05	26.32	62.98	70.83
50	75.18	98.58	60.91	42.25	25.74	58.86	70.83
60	74.33	98.49	58.43	43.17	25.29	37.31	70.83
70	73.44	98.36	56.13	44.13	24.12	41.76	70.83
80	72.91	98.26	51.27	42.86	22.79	41.76	70.83
90	72.22	97.86	49.07	42.83	23.97	41.76	70.83
100	71.11	97.65	47.58	44.40	22.50	41.76	70.83
150	67.43	97.00	35.14	43.77	19.85	41.76	70.83
200	62.32	96.00	27.57	40.67	16.47	41.76	70.83
250	57.50	94.65	21.01	53.76	12.79	41.76	70.83

Table 4: Classification accuracies (%) on RP

Dim.(k)	IRMA		COIL-100		ZuBuD		MIAS		Outex		Colon		Leukemia	
	Acc.	SD	Acc.	SD	Acc.	SD	Acc.	SD	Acc.	SD	Acc.	SD	Acc.	SD
5	30.46	2.77	63.24	2.69	3.77	0.86	44.05	2.65	17.97	2.01	66.83	8.39	70.44	11.10
10	42.14	3.62	90.54	0.71	9.64	1.66	43.03	2.50	20.37	1.39	72.37	5.06	77.19	7.70
15	47.86	3.93	94.95	0.50	15.46	1.63	42.60	2.41	21.04	1.51	75.75	5.69	81.75	5.52
20	50.89	3.99	96.57	0.35	20.66	1.70	41.93	1.81	20.96	1.25	76.94	5.32	83.25	7.64
25	53.19	3.75	97.26	0.28	25.28	1.33	42.00	2.00	20.82	1.33	77.15	5.82	82.37	6.08
30	54.89	3.57	97.62	0.28	28.69	1.53	41.88	2.27	20.62	1.20	78.01	4.97	84.74	5.01
35	55.71	3.73	97.83	0.24	31.83	1.54	42.28	2.38	20.67	1.27	78.01	4.38	83.95	5.98
40	56.72	3.62	98.01	0.21	34.08	1.55	42.04	2.35	20.75	1.16	77.85	4.82	85.44	4.98
45	57.61	3.18	98.14	0.19	36.23	1.67	42.16	2.40	20.43	0.77	77.96	4.09	85.70	5.21
50	58.03	2.93	98.22	0.17	37.99	1.52	42.02	2.28	20.34	0.93	78.49	3.64	86.84	5.39
60	59.25	2.75	98.37	0.15	40.80	1.54	42.07	2.12	20.30	0.93	78.76	3.23	87.89	5.07
70	59.96	2.45	98.46	0.16	43.15	1.24	41.97	2.02	20.03	0.93	79.19	2.98	87.81	4.63
80	60.73	2.37	98.50	0.14	44.78	1.46	41.77	1.77	20.19	0.86	79.30	2.70	88.60	4.57
90	61.27	2.06	98.56	0.10	46.27	1.60	41.75	1.72	20.17	0.85	79.68	2.93	89.74	3.42
100	61.98	2.08	98.59	0.11	47.48	1.63	41.67	1.87	20.12	0.76	79.89	3.22	90.09	3.94
150	64.35	1.68	98.69	0.10	50.86	1.44	41.50	1.92	20.26	0.68	80.22	2.63	91.32	3.79
200	66.00	1.09	98.75	0.10	52.80	1.25	41.41	1.36	20.04	0.67	80.22	2.70	90.70	4.12
250	67.01	0.90	98.79	0.08	54.01	1.01	41.38	1.42	20.09	0.62	80.16	2.43	90.53	3.37

Table 5: Average testing time (in seconds)

Dim. (k)	IRMA	COIL-100	ZuBuD	MIAS	Outex	Colon	Leukemia
5	71	46	0.93	0.11	0.44	0.01	0.01
10	137	87	1.76	0.20	0.81	0.02	0.01
15	207	129	2.57	0.27	1.17	0.03	0.01
20	278	172	3.39	0.37	1.56	0.02	0.02
25	344	216	4.20	0.45	1.94	0.02	0.01
30	404	258	5.01	0.53	2.31	0.03	0.01
35	478	339	5.84	0.61	2.69	0.03	0.01
40	541	344	6.65	0.70	3.08	0.03	0.01
45	609	388	7.44	0.78	3.57	0.04	0.01
50	676	433	8.25	0.87	3.86	0.04	0.01
60	809	517	9.91	1.05	4.55	0.05	0.02
70	941	617	11.57	1.23	5.44	0.05	0.02
80	1073	698	13.20	1.38	6.48	0.06	0.03
90	1206	770	14.82	1.56	6.79	0.06	0.03
100	1429	855	16.63	1.76	7.59	0.07	0.03
150	1998	1275	24.60	2.55	11.36	0.10	0.04
200	3279	1698	32.87	3.45	15.02	0.14	0.05
250	3354	2175	41.47	4.33	18.75	0.17	0.07
All	13399	8618	168.23	15.50	72.70	1.29	1.77