1 APPENDIX: DATA USED IN THIS STUDY

All the data used in this study is available at http://promisedata.org/data or from the authors. Our data includes:

- Data from the International Software Benchmarking Standards Group (ISBSG);
- The Desharnais and Albrecht data sets;
- SDR, which is data from projects of various software companies from Turkey. SDR is collected from Softlab, the Bogazici University Software Engineering Research Laboratory repository [1];
- And the standard COCOMO data sets (Cocomo*, Nasa*).

Projects in ISBSG dataset can be grouped according to their business domains. In previous studies, breakdown of ISBSG according to business domain has also been used [1]. Among different business domains we selected banking due to:

1. Banking domain includes many projects whose data quality is reported to be high (ISBSG contains projects with missing attribute values).
2. ISBSG Banking domain is the dataset we have analyzed and worked for a long time due to our hands on experience in building effort estimation models in banking industry.

We denote the banking domain subset of ISBSG as "ISBSG-Banking".

Note that two of these data sets (Nasa93c2, Nasa93c5) come from different development centers around the United States. Another two of these data sets (Cocomo81e, Cocomo81o) represent different kinds of projects:

- The Cocomo81e "embedded projects" are those developed within tight constraints (hardware, software, operational, ...);
- The Cocomo81o "organic projects" come from small teams with good experience of working with less rigid requirements.

The skewness of our effort values (2.0 to 4.4): our datasets are extremely heterogeneous with as much as 40-fold variation. There is also some divergence in the features used to describe our data:

- While our data includes some effort value (measured in terms of months or hours), no other feature is shared by all data sets.
- The Cocomo* and NASA* data sets all use the features defined by Boehm [2]; e.g. analyst capability, required software reliability, memory constraints, and use of software tools.
- The other data sets use a wide variety of features including, number of entities in the data model, number of basic logical transactions, query count and number of distinct business units serviced.

REFERENCES

[1] A. Bakir, B. Turhan, and A. Bener, "A new perspective on data homogeneity in software cost estimation: A study in the embedded systems domain," *Software Quality Journal*, 2009. [Online]. Available: http://bit.ly/fp4b9N

[2] B. W. Boehm, *Software Engineering Economics*. Prentice Hall PTR, 1981.