RINGKASAN

PENERAPAN METODE *SQUASHING* PADA DATA BERUKURAN BESAR YANG BERDISTRIBUSI PARETO TERGENERALISIR, Rimuljo Hendradi, Eto Wuryanto dan Dyah Herawatie, 2005, 30 halaman.

Data *warehouse* (data yang berukuran sangat besar) merupakan hasil dari pesatnya perkembangan teknologi informasi akibat penggunaan database. Dengan jumlah data yang sangat besar maka baik analisis maupun visualisasi data dengan metode-metode tradisional sulit dikerjakan bahkan tidak bisa dilakukan. Sehingga perlu dicarikan cara supaya dapat dilakukan pengolahan data yang berukuran besar tersebut. Cara paling sederhana adalah melakukan pengurangan jumlah data (reduksi data) dengan metode sampling tradisional (konvensional), misalkan metode sampling acak sederhana, sistematik dan stratifikasi. Tetapi cara ini masih diragukan apakah data hasil sampling tersebut sudah mewakili *massive dataset* yang merupakan data induknya.

Penelitian terbaru yang dilakukan DuMouchel et al (1999) mengusulkan suatu pendekatan baru yang disebut metode *squashing,* yang mereduksi *massive dataset* menjadi *dataset* yang lebih kecil dan dapat mempresentasikan data induk. Dalam reduksi data, sifat data dengan "*heavy tail*" merupakan hal yang sangat menarik karena sampel yang dihasilkan dari jenis data ini diperlukan jumlah yang besar supaya dapat mempresentasikan *dataset* induk. Salah satu data dengan "*heavy tail*" ini adalah data yang berdistribusi Pareto tergeneralisir yang diindikasikan oleh parameter *tail*-nya.

Penelitian ini bertujuan memperoleh sampel dari data berukuran besar yang berdistribusi Pareto tergeneneralisir, yang dapat diolah secara statistik dengan mudah dan menghasilkan akurasi sesuai keinginan dalam proses pengambilan keputusan (inferensi). Untuk tujuan ini akan dibandingkan sampling hasil dari metode sampling tradisional (metode

sampling acak sederhana, sistematik, dan stratifikasi) dan metode *squashing,* dengan menggunakan indikator nilai MSE.

Untuk mencapai tujuan di atas digunakan metode penelitian berikut : penyusunan algoritma yang meliputi antara lain membangkitkan data yang berdistribusi Pareto tergeneralisir; penentuan sampel dengan menggunakan metode tradisional dan metode *squashing*; estimasi parameter dari Pareto tergeneralisir untuk data *squashing* dan non data *squashing* dengan menggunakan metode maksimum *likelihood* dan Newton-Raphson. Selanjutnya, algoritma tersebut disusun ke dalam program komputer (dengan software S-plus).

Reduksi data atau pembuatan sampel dengan metode *squashing* dapat dilakukan dengan cara : Pertama, melakukan pengelompokkan terhadap data induk ke dalam beberapa partisi atau kelompok yang sama. Kedua, untuk per kelompok jumlah anggotanya berfungsi sebagai nilai pembobot dan nilai *pseudo point*-nya sama dengan rata-rata dari data di masing-masing kelompok.

Penerapan metode *squashing* dalam penentuan sampel dari data induk yang beristribusi Pareto Tergeneralisir bisa dilakukan dengan cara : Pertama, melakukan pengelompokkan terhadap data induk ke dalam beberapa kelompok yang sama. Kedua, pada tiap kelompok dihitung secara acak r ($r > 1$) nilai pembobot dan r nilai *pseudo point.*

Dari data hasil simulasi, untuk metode tradisional yang cenderung memberikan hasil estimasi yang lebih baik adalah metode sampling sistematik dan metode stratifikasi. Jika dibandingkan dengan metode tradisional, metode squashing menunjukkan hasil yang lebih baik. Hal ini diindikasikan dengan nilai MSE yang lebih kecil untuk estimator dari α dan β.

(Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Airlangga, No. Kontrak 729/J03.2/PG/2005 Ditjen Dikti, Depdiknas)

SUMMARY

APPLICATION OF SQUASHING METHODS TO THE LARGE DATA FROM GENERALIZED PARETO DISTRIBUTION, Rimuljo Hendradi, Eto Wuryanto dan Dyah Herawatie, 2005, 30 pages.

The database use in information technology that growth very fast effect the presentation of data warehouse (the very large data). Either data analysis or visualization is difficult to be worked using traditional method for the data warehouse even cannot be done. So it is necessary to obtain the methods that can analyse this data type. The simpiest way is to reduce the number of data (data reduction) using the traditional (conventional) sampling method : simple random sampling, systematic and stratification. But this way is doubt what the data sampling have represented the massive dataset that is the main data.

The newest research that is done by Dumouchel et al (1999) propose a new approach so-called squashing method. This methods reduce the massive dataset become the smaller dataset and can present the main data. In data reduction context, heavy tailed data are interesting because the sample that is yielded from this data type need a large size so can present the main dataset. One of heavy tailed data is data from generalized Pareto distribution which indication by its tail parameter.

This research aim to obtain the sample from the large data of generalized Pareto distributed that can be processed statistically easily and yield the certain accuration in decision making. The purpose will be reach by compare a sample of traditional sampling method (simple random sampling, systematic and stratification) and one of squashing method using indicator MSE.

The following is the research method that is used are construct any algorithm : generate data from generalized Pareto distribution; determination of either a sample of traditional method or one of squashing method; estimate the parameter of generalized Pareto distribution for

squashing data and non squashing data by applying maximum likelihood method and Newton-Raphson method. Then the algorithm is compiled into computer program (in S-Plus programming).

In squashing method, data reduction or making sample can be done by : First, grouping the main data into some same partition. Second, each group the sum of member as a weighted value and the pseudo point equal to the mean of data.

Application of squashing method in determination of sample from the main data of generalized Pareto distributed can be conducted by : First, cluster the main data into some same group. Second, each group calculate randomly r ($r > 1$) weighted value and r pseudo point.

By doing to simulation data can is obtained, for the traditional method, systematic sampling method and stratification method tend to give a better estimation. In fact squashing method have the result better than traditional method because the MSE value of the α and β estimator using squashing method is smaller.

(Department of Mathematics, Faculty of Mathematics and Natural Sciences Airlangga University, The Contract Number 729/J03.2/PG/2005 Ditjen Dikti, Depdiknas)