From Census to Grids: Comparing Gridded Population of the World with Swedish Census Records

Ola Hall[*,1], Emilie Stroh[2] and Fredy Paya[3]

[1]*Department of Human and Economic Geography, Lund University, Sweden*

[2]*Department of Occupational and Environmental Hygiene. Lund University, Sweden*

[3]*University of Waterloo, Canada*

Abstract: The increased availability of digital spatial data combined with improved capabilities of Geographic Information Systems (GIS) have allowed for the development of several global population distribution databases, such as the GPW, and LandScan. Making population distribution data available as a high-resolution raster database which facilitates rapid GIS analysis at the local level and for any zoning. Due to the complex nature of population as a geographical variable, several approaches have been adopted to estimate their spatial distribution, including statistical modeling, surface modeling, and cartographic methods. However, many of these methods require assumptions that over-simplify the reality or disaggregate population totals based on the heuristic or empirical parameters. Recently, critical voices were heard, questioning the quality and usability of gridded population data.

In this paper, we compare gridded population data products for parts of Sweden with high-resolution population records obtained from the Swedish National Registry through the Regional Office of Scania, Sweden. Ground-truth consists of the total population in Scania located as points at the center coordinates of their real estate (located by the Swedish Land Survey). Results indicate that there are significant differences between compared datasets.

Keywords: Population, GIS, census, LandScan, ground, EU+27.

1. INTRODUCTION

Efforts to estimate population distribution for a regular raster grid predate the computerization of geography that started in the 1980s [1]. Early examples such as the map by Adams [2] for West Africa served largely the cartographic purposes. Census offices, most notably those of Japan and Sweden, also produced national population grids for inclusion in the national atlases. Computerized population maps for individual countries were produced by the US Census Bureau using rectangular grid cells superimposed with circles for major urban areas [3]. Deichmann and Eklund hpresented a continental, gridded population database for Africa which was used to investigate interactions between population and land degradation [4]. Others, such as Martin and Bracken, developed techniques for producing local-level population grids [5].

Applications are found in various fields, from disaster management, hazards and vulnerability science, climate change science to human welfare and public health. Few studies in demography make use of this kind of data. In fact, critical voices regarding quality have been raised [6]. It seems like they are mostly used in regions with poor census records. Usually one cannot reliably assess how accurate the resulting distributions are because there is no basis for sound validation [7].

*Address correspondence to this author at the Department of Human and Economic Geography, Lund University, Sweden; Tel: 46 (0) 46-2220000; E-mail: ola.hall@keg.lu.se

The aim of this study is to investigate the quality of four gridded population datasets, Gridded Population of the World (GPW), LandScan, Global Rural-Urban Mapping Project (GRUMP) and a recent dataset covering the European community (EU+27).

2. DATA AND METHODS

2.1. Ground Truth

Due to the structure and frequent use of personal ID numbers in Sweden the possibility of combining data at individual level from different sources is substantially better than many other countries. For example, by combining an individual's ID number with the Swedish National Land Survey's property registers the coordinates of and information on an individual's residential address can be linked to that person. The census data used in this study were obtained from the Swedish National Registry through the Regional Office of Scania, Sweden. The dataset referred to as "A" consists of the total population in Scania on the 31st of December in 2001, i.e. 1,129,059 individuals located as points at the center coordinates of their real estate (located by the Swedish Land Survey).

2.2. Population Density Grid of EU-27+

This dataset covers the 27 member states of EU and Croatia. The dataset is referred to as "dataset B" in the rest of the text. The spatial resolution is 100 x 100 m (1 ha) and in projected in Lambert-Azimuthal equal area projection. Values in the grid correspond to population density and to obtain the population count the sum of pixel values are

1874-9232/12

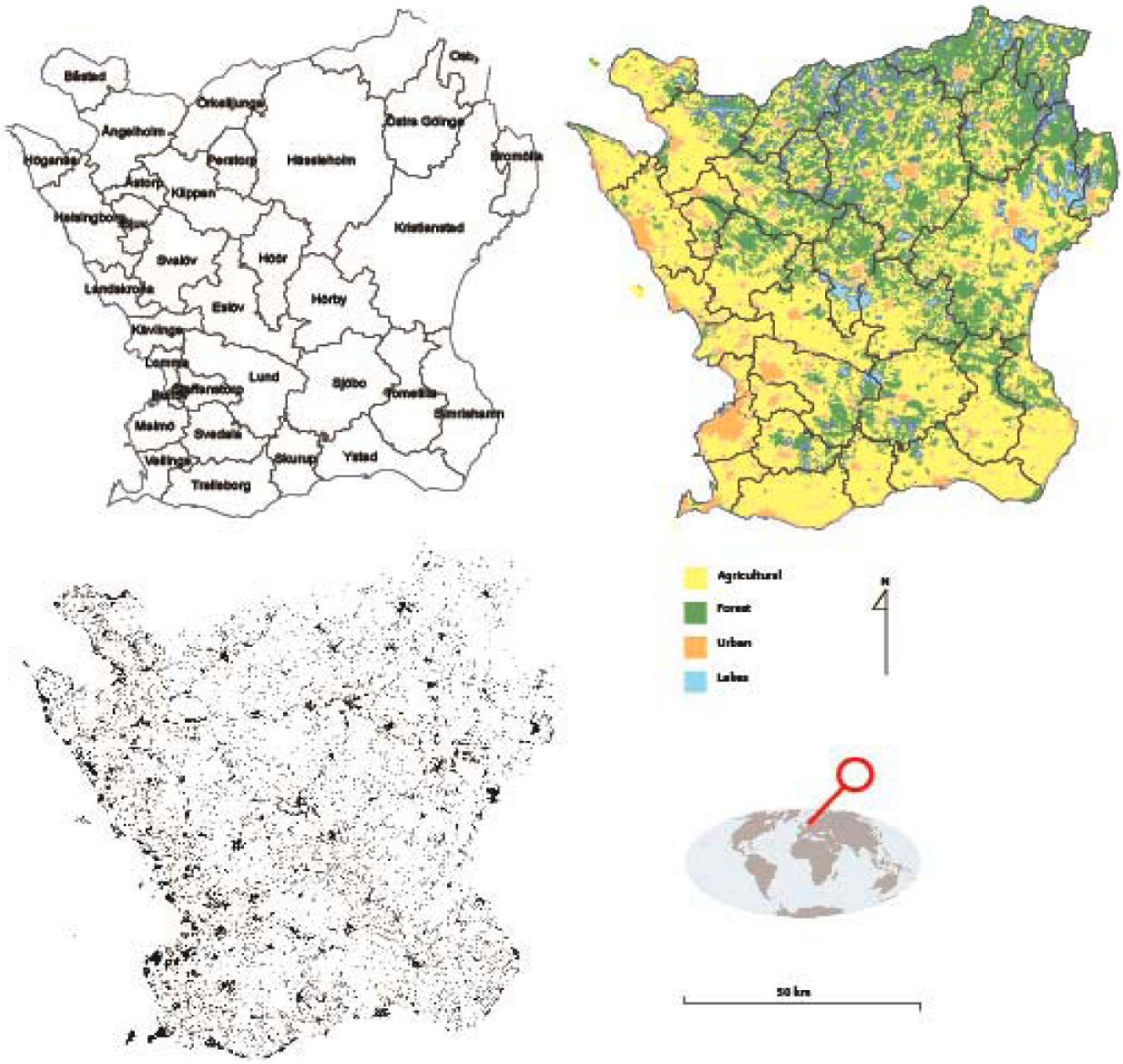

Fig. (1). Top left: Study area and name of communes. Top: right: Simplified land cover map of the Skåne region. Lower left: population distribution based SCB data.

divided by 100. Input data is population by commune (census 2001) provided by Eurostat.

The gridded values are obtained from a disaggregation scheme with CORINE Land Cover (CLC-2000) and reclassified into 9 classes (Table **1**).

Additional data were LUCAS-2001 point data provided by Eurostat and population density grids from 2006 (1 km spatial resolution) provided by National Statistics Institutes members of the European Forum for Geostatistics.

The model used assumes that population density can be expressed as

$$Y_{cm} = U_{ch}W_m$$

where Y_{cm} is the density of population for land cover type, c in commune and m that belongs to stratum h. This implies two simplifying assumptions A) the population density is supposed to be the same for all pixels in the same commune and same CLC class and b) the ratio between the population density of two land cover classes is constant for all communes in the stratum. U_{ch} are coefficients. A full description is found in Gallego [8]. The full dataset is downloadable from http://dataservice.eea.europa.eu/dataservice/.

2.3. Gridded Population of the World and Global Urban Mapping Project

Gridded Population of the World, Version 3 (GPWv3) consists of estimates of human population for the years 1990, 1995, and 2000 by 2.5 arc-minute grid cells and associated

datasets dated circa 2000. The data products include population count grids (raw counts), population density grids (per square km), land area grids (actual area net of ice and water), mean administrative unit area grids, centroids, a national identifier grid, national boundaries, and coastlines. For this project population count grids for 2001 were used and this was referred to as "dataset C".

Table 1. CLC Aggregation Scheme

Grouped Class	CORINE Class	Label
1	111	Continuous urban fabric
2	112	Discontinuous urban fabric
3	121,133,14	Other urban
4	122-124, 131-132	Low population artificial
5	21,22,23	Agriculture
6	241-243	Heterogeneous
7	244,31	Forest
8	32	Natural vegetation
9	33,4,5	Bare land, wetland and water

The GPW database uses two basic inputs: non-spatial population estimates (i.e. tables of population counts listed by area names) and spatially explicit administrative boundary data. A proportional allocation gridding algorithm (areal weighting scheme), utilizing more than 300,000 national and sub-national administrative units, is used to assign population values to grid cells. Dataset C is produced by the Columbia University Center for International Earth Science Information Network (CIESIN) in collaboration with Centro Internacional de Agricultura Tropical (CIAT).

The allocation mechanism for the Global Rural Urban Mapping Project (GRUMP) builds on the GPW approach but explicitly considers the population of urban areas. The GRUMP, Alpha Version consists of estimates of human population for the years 1990, 1995, and 2000 by 30 arc-second grid cells and associated datasets dated circa 2000. The data products include population count grids (raw counts), population density grids (per square km), land area grids (actual area net of ice and water), mean geographic unit area grids, urban extent grids, centroids, a national identifier grid, national boundaries, coastlines, and settlement points A proportional allocation gridding algorithm, utilizing more than 1,000,000 national and sub-national geographic units, is used to assign population values to grid cells. For this project GRUMP population counts for year 2000 was used and is henceforth referred to as "dataset D".

2.4. LandScan

At approximately 1 km resolution (30 arc-sec), LandScan is the finest resolution global population distribution data available and represents an ambient population. The LandScan spatial data and imagery analysis technologies and a multi-variable dasymetric modelling approach were used to disaggregate census counts within an administrative boundary. Since no single population distribution model can account for the differences in spatial data availability, quality, scale, and accuracy as well as the differences in cultural settlement practices, LandScan population distribution models are tailored to match the data conditions and geographical nature of each individual country and region. This dataset is referred to as "dataset E".

The values of the cells are population counts representing an average, or ambient, population distribution. An ambient population integrates diurnal movements and collective travel habits into a single measure [9]. Since natural or man-made emergencies may occur at any time of the day, the goal of the LandScan model is to develop a population distribution surface in total, not just the locations of where people sleep.

The dataset has a spatial resolution of 30 arc-seconds and is output in a geographical coordinate system - World Geodetic System (WGS) 84 datum. The 30 arc-second cell, represents 1 km^2 near the equator. The values of the cells are integer population counts, not population density, since the cells vary in size. Population counts are normalized to sum to each sub-national administrative unit estimate. For this reason, projecting the data in a raster format to a different coordinate system (including on-the-fly projections) will result in a re-sampling of the data and the integrity of normalized population counts will be compromised.

2.5. Experimental Design

The experimental design for this paper is straightforward and as follows. For each of the dataset (B-E) grids were created with the spatial dimensions replicated. Dataset B-F-was cropped to fit the extent of region Skåne (Fig. **1**). Each newly generated grid was updated with the point data from dataset A. Dataset B-E was then compared to up-scaled versions of dataset A. Global statistics including correlation coefficients were calculated and summarized in table 2.

A difference dataset for each combination of A and B-E was computed using the following formula

$$D = (A - B) / (B + A)$$

where A is a version of dataset A and B is dataset B-E. The resulting index ranges from -1 to +1 with 1 indicating an overestimation of population and -1 indicating an underestimation of population. A value around zero indicates similarity between datasets. Resulting difference maps are found in Fig. (**2**).

3. RESULTS

Global statistics reveal differences among dataset B-E in estimating population in the Skåne region (Table **2**). In terms of statistical correlation, dataset E outperforms the rest (r^2=0.59) followed by dataset C (r^2=0.34). Dataset B replicates the statistical properties (average) fairly well. In terms of estimating the total population in the whole region dataset C and E show the best performance. It should be kept in mind that there is a slight difference in the timing of censuses and data used among all datasets that can account for some of the differences.

The general impression from the difference maps is that the precision of population estimates is rather poor. Dataset C and D overestimate population in cities and underestimate in the transition zone between urban and rural. While the performance of dataset C is generally poor and

Table 2. Global Statistics for Dataset A-E

	SCB (A)	Landscan (E)	EU (B)	GPW (C)	GRUMP(D)
Max	15	8467	95	20281	814
Min	0	0	0	1	0
Average	1.0	65	0.8	1106	55
Sum	1,129,059	1,048,357	847,893	1,136,876	899,176
Resolution	~0.1 km^2	~1 km^2	~0.1 km^2	~20 km^2	~1 km^2
R^2	Na	0.59	0.21	0.34	0.18
N	1048592	16054	1048592	1029	16226

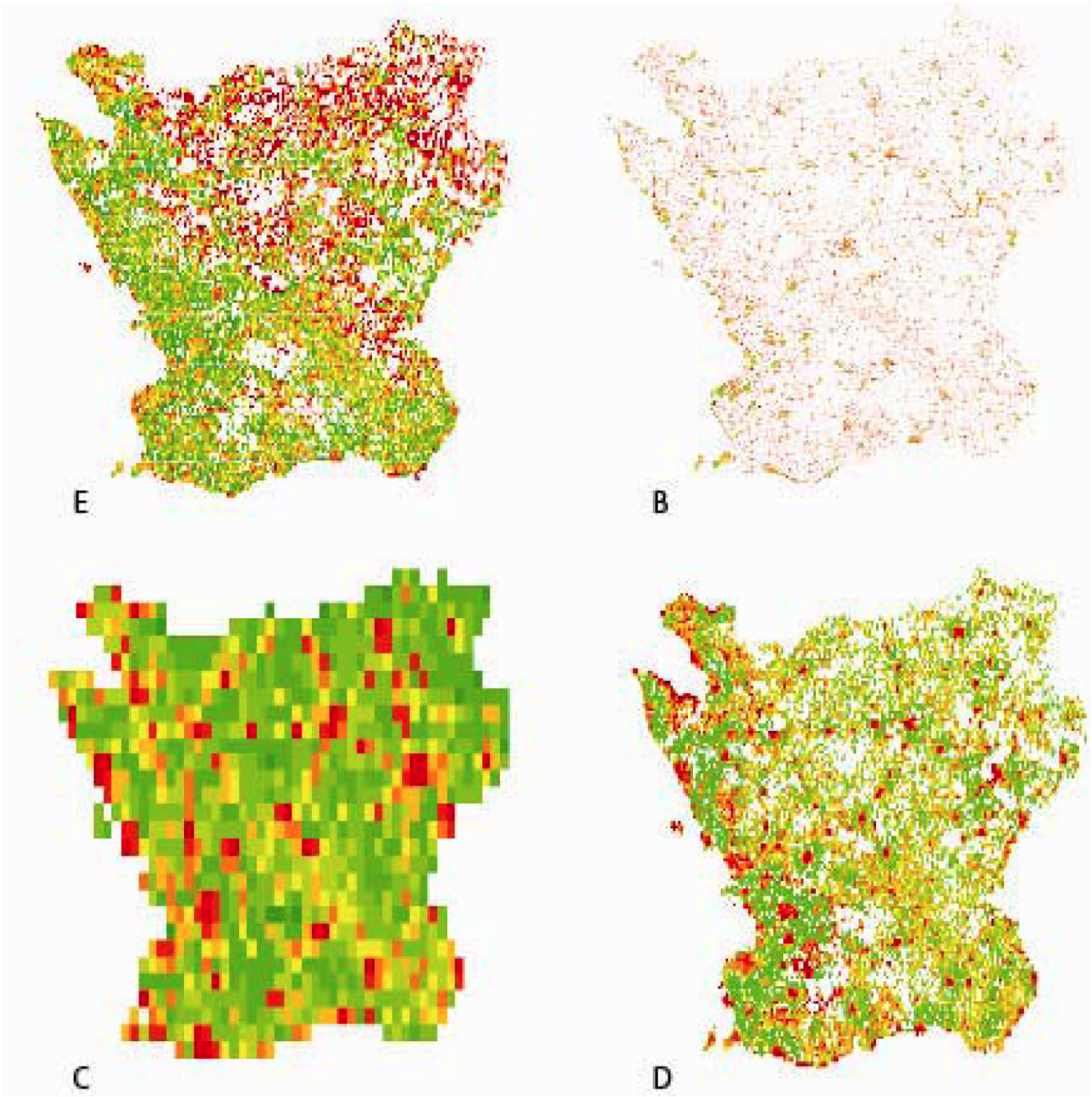

Fig. (2). Difference maps. B=EU 27+, C= GPWv3, D= GRUMP and E= Landscan.

predominantly underestimates population, dataset D performs rather well transcending away from the west coast. The high-resolution dataset B overestimates population heavily. In dataset E, a southwest-northeast gradient pattern is visible. In the north population is overestimated while in the south it is underestimated. This pattern corresponds well with the pattern of the forest region and the agricultural region in Skåne (cf. Fig. **1**).

CONCLUSIONS

The aim of this study was to investigate four gridded population databases. The rationale behind this study was that these types of population databases are being increasingly used and very few evaluations of quality exist. Based on this a simple experiment was conducted based on the availability of high-resolution population data for Skåne, Sweden, acting as a reference dataset.

The general conclusion is that big differences exist among gridded population databases as well as between reference data and gridded population databases. In terms of global statistics, Landscan data (dataset E) seems to perform best. It is quite a surprise that some of the datasets have problems reporting correct total population numbers (e.g. B and D). When we look at the difference maps a somewhat different pattern emerges. Dataset D has large regions where the estimated population count is near the values of reference data. Major differences are found close to the larger urban areas in the southwest and surroundings. We have also noted a regional pattern in dataset E that divides the study region in two halves. We cannot provide an explanation for this observed pattern but can conclude that the pattern corresponds with the extent of the forest- and agricultural regions of Skåne.

It has been argued that the value of gridded population datasets can be in disaster relief efforts. It has been used to estimate population at risk [10]. In disaster relief efforts it is important that resources and time are not diverted from the affected population [11]. One risk with population mapping at this high spatial resolution is that people are assigned by the model to cells where actually no one lives. All datasets except Landscan seem to have a problem with this. The main use of this type of data is in developing countries with poor data infrastructure. The relative benefits are manifold compared to the region of Skåne where excellent population records exist. Still, there is a need for more research on the quality of gridded population data with reference data from target nations.

ACKNOWLEDGEMENT

None declared.

CONFLICT OF INTEREST

None declared.

REFERENCES

[1] Freire S. Modelling of spatiotemporal distribution of urban population at high resolution – value for risk assessment and emergency management. Konecny M, *et al*., Eds. Geographic information and cartography for risk and crisis management, lecture notes in geoinformation and cartography, Springer-Verlag Berlin Heidelberg 2010; DOI 10.1007/978-3-642-03442-8_4.

[2] Adams JA. Population map of West Africa. graduate school of geography. In: Determining global population distribution: methods, applications and data. London: London School of Economics. Adv Parasitol, Discussion Paper No. 26, 1968.

[3] Leddy R. Small area populations for the United States. Paper presented at the association of American geographers annual meeting in San Francisco. Geographic Studies Branch, International Programs Center, US Bureau of the Census, Washington, DC 1968.

[4] Deichmann U, Eklundh L. Global digital datasets for land degradation studies: a GIS approach. Nairobi, Kenya: United Nations environment programme, global resource information database, case study No. 4, 1994.

[5] Clarke JI, Rhind DW. Population data and global environmental change. Human dimensions of global environmental change programme report 3. New York: international social science council, 1992.

[6] Jordan L. Eyes from above: remote sensing and virtual globes. In spatial demography meeting of the population association of America, New York 1997.

[7] Balk DL, Deichmann U, Yetman G, Pozzi F, Hay SI, Nelson A. Determining global population distribution: methods, applications and data, Adv Parasitol 2006; 62: 119-56.

[8] Gallego FJ. A population density grid of the European Union. Popul Environ 2010; 31: 460-73.

[9] Dobson JE, Bright EA, Coleman PR, Durfee RC, Worley BA. LandScan: a global population database for estimating populations at risk. Photogrammetric Eng Remote Sens 2000; 66: 849-57.

[10] Hall O, Duit A, Caballero L. World poverty, environmental vulnerability and population at risk for natural hazards. J Maps 2008; pp. 151-60.

[11] Harvard Humanitatian Initiative. Disaster Relief 2.0: The future of information sharing in humanitarian emergencies. Washington, D.C. and Berkshire, UK: UN Foundation & Vodafone Foundation Technology Partnership 2011.

Received: September 26, 2011 Revised: October 18, 2011 Accepted: October 22, 2011