Jurnal Teknologi, 39(B) Dis. 2003: 47–62

THE EFFECTS OF DATA AGGREGATION ON THE SPATIAL ANALYSIS OF POPULATION CONCENTRATION

RUSLAN RAINIS[1] & NORESAH MOHD SHARIFF[2]

Abstract. This article describes a study on the effects of data aggregation on the measure of spatial concentration of population distribution for Peninsular Malaysia. Spatial concentration of population is measured using Hoover population concentration index. Census data at the district level from 1980 to 2000 was used as the base for the analysis. Analysis of data aggregation was carried out using three different data levels: district, state and region. This study found that in general, data aggregation affects the measure of spatial concentration of population distribution. The more aggregated the data is, the lower the measure of population concentration, and vice versa. More importantly, the study also found that data aggregation might not necessarily replicate the trend of population concentration at the detailed data level. Using the district level data, the population distribution of Peninsular Malaysia during the twenty years period was neither dispersed nor concentrated but relatively uniform. The population distribution was slightly more dispersed in 1991 but became more concentrated in year 2000 as compared to 1980, with concentration indices changed from 49.24 to 48.85, and then 51.57. However, as the data were aggregated to the regional and state levels, the population distribution becomes more dispersed. The trend of population concentration at the regional level was similar to that of the district data. Interestingly, when using the state data, the trend of population concentration is different, whereby the population distribution continuously becomes more concentrated. Similar pattern was observed at the regional level. With such variations in results, it is recommended that the appropriate data aggregation is used in any study pertaining to the spatial concentration of population.

Keywords: Data aggregation, population concentration, hoover index, geographical information system (GIS)

Abstrak. Artikel ini menerangkan satu kajian kesan agregasi data pada ukuran penumpuan ruangan taburan penduduk bagi Semenanjung Malaysia. Penumpuan ruangan penduduk diukur menggunakan indeks penumpuan penduduk Hoover. Data banci di peringkat daerah dari tahun 1980 hingga 2000 telah digunakan sebagai asas kepada analisis yang dijalankan. Analisis agregasi data telah dijalankan menggunakan tiga peringkat data iaitu daerah, negeri dan wilayah. Secara umum, kajian ini mendapati agregasi data memberi kesan kepada pengukuran penumpuan ruangan taburan penduduk. Ukuran penumpuan penduduk menjadi semakin kecil apabila data semakin diagregatkan dan sebaliknya. Yang lebih penting kajian ini juga mendapati agregasi data tidak semestinya mengulangi tren penumpuan penduduk pada peringkat data yang lebih terperinci. Dengan menggunakan data di peringkat daerah, taburan penduduk Semenanjung Malaysia bagi jangkamasa 20 tahun kajian adalah secara relatifnya seragam iaitu tidak berselerak mahupun bertumpu. Taburan penduduk telah sedikit berselerak pada tahun 1991 tetapi menjadi semakin bertumpu pada tahun 2000 berbanding tahun

[1] Coordinator, Geoinformatic Unit, Geography Section, School of Humanities, Universiti Sains Malaysia, 11800 Penang. e-mail: rruslan@usm.my. Corresponding author

[2] Geography Section, School of Distance Education, Universiti Sains Malaysia, 11800 Penang. e-mail: noresahms@yahoo.com. (on study leave at University of Manchester, UK)

1980 yang mana indeks penumpuan penduduk telah berubah dari 49.24 ke 48.85, dan kemudiannya 51.57. Bagaimanapun, taburan penduduk semakin berselerak apabila data diagregatkan ke peringkat negeri dan wilayah. Tren penumpuan penduduk di peringkat wilayah adalah sama dengan peringkat daerah. Yang menarik ialah apabila data peringkat negeri digunakan, tren penumpuan penduduk adalah berbeza yang mana taburan penduduk menjadi lebih bertumpu secara berterusan. Corak yang sama telah diamati pada peringkat wilayah. Dengan variasi yang sedemikian, maka adalah dicadangkan supaya agregasi data yang sesuai digunakan dalam sebarang kajian berkaitan penumpuan ruangan penduduk.

Kata kunci: Agregasi data, penumpuan penduduk, indeks hoover, sistem maklumat geografi (GIS)

1.0 INTRODUCTION

Studies on the spatial patterns of population distribution are very important. The concentration of economic development and of population are, in fact, interrelated: as a country's economic development concentrates in the core (usually the national capital), so does its population (Portnov & Pearlmutter, 1999). On the other hand, sufficient population size is required to support any development programs especially in terms of labor supply. If this phenomenon continues unchecked, the severity of socio-economic inequalities and imbalances between different regions will increase, a challenge faced by many countries. As the spatial distribution of a nation's or state's population is inextricably linked to changes in its socio-economic and political organization, the changes in spatial distribution of population can be an important indicator in monitoring the performance of the various socio-economic policies implemented by the government. This information is important in the planning process and formulation of policy.

There have been numerous studies on population concentration and dispersal for various countries of the world especially the United States (eg. Hoover, 1941; Lichter, 1985; Otterstrom, 2001; Vining & Strauss, 1977; Souza, 2001). Many methods could be used to measure the distribution of population (Hoover, 1941; Duncan, 1957). These include the Hoover index, entropy index and Gini index. As population data is collected based on areal spatial unit, the level of spatial aggregation might affects the results from using such methods. However, to date little study has been carried out to determine the sensitivity of data aggregation upon the uses of such methods especially in the context of Malaysia. This kind of analysis is important because data is usually collected at a specific areal unit (such as mukim or district), but might be used for various purposes at much broader level of spatial unit (such as state, region or national). With the advancement of geographical data handling systems such as geographical information system (GIS), data at the highest detail could be managed and manipulated quite easily.

The objective of this article is to report the results of a study on the effects of data aggregation on the measure of population concentration using data for Peninsular Malaysia between 1980 and 2000 as the case study. It is hoped that the study will fill in the gap and contributes to the body of knowledge on the geography of population

distribution especially in the context of Malaysia. It is hoped that such information will be useful to planners and decision makers as a guide to the selection of appropriate data level in the study of population distribution.

2.0 DATA AND METHODS

There are a number of methods to measure population concentration such as the Hoover index (Hoover, 1941), entropy index and Gini index. The suitability of each method greatly depends on the purpose and availability of data. However, in the study of population distribution, Lichter (1985) suggested that Hoover concentration index (Hoover, 1941) as employed by Duncan *et al.* (1961) and Vining and Strauss (1977) can provide a useful measure of national patterns of population concentration. In the present study, the Hoover index is employed to examine the trends of population concentration and dispersion at various scales. This index is a timeworn measure which gives an easily comparable, relative value of concentration among various sizes of geographic units (Otterstrom, 2001). This measure, H_t can be calculated in the following manner:

$$H_t = \tfrac{1}{2} \sum_{i=1}^{k} \left| p_{it} - a_i \right| \qquad (1)$$

where p_{it} is percentage of a nation's population in district i at time t, a_i represents the percentage of the nation's land area covered by district i and k is the total number of districts. If p_{it} is equal to a_i for all districts, then population spread over all the districts in proportion to land area and H_t is equal to 0. This indicates a perfectly dispersed pattern of population distribution. The distribution of population across districts becomes increasingly concentrated as H_t approaches 100. Neither extreme is likely, but the relative change in the value over time can be used to track spatial changes in the population. The increase in value of H_t with time indicates a pattern of increasing population concentration with time, whilst a decreasing H_t with time suggests the dispersal or deconcentration of population.

This study is based on the population data as reported in the various Census of Population reports. The latest census for Malaysia was carried out in year 2000 and a few preliminary results have been published by the Malaysian Department of Statistics since 2001. So far, the most detailed population data for the Census 2000 is only available at the district level (Figure 1 and Table 1), hence, this study uses district as the basic spatial unit. The district is the smallest geographic unit with the most stable boundary and does not change very much between censuses. In the 2000 census, Peninsular Malaysia was divided into 82 districts, an addition of one district as compared to the 1991 census. To ensure compatibility between censuses, this study uses the 1991 census boundary as the base.

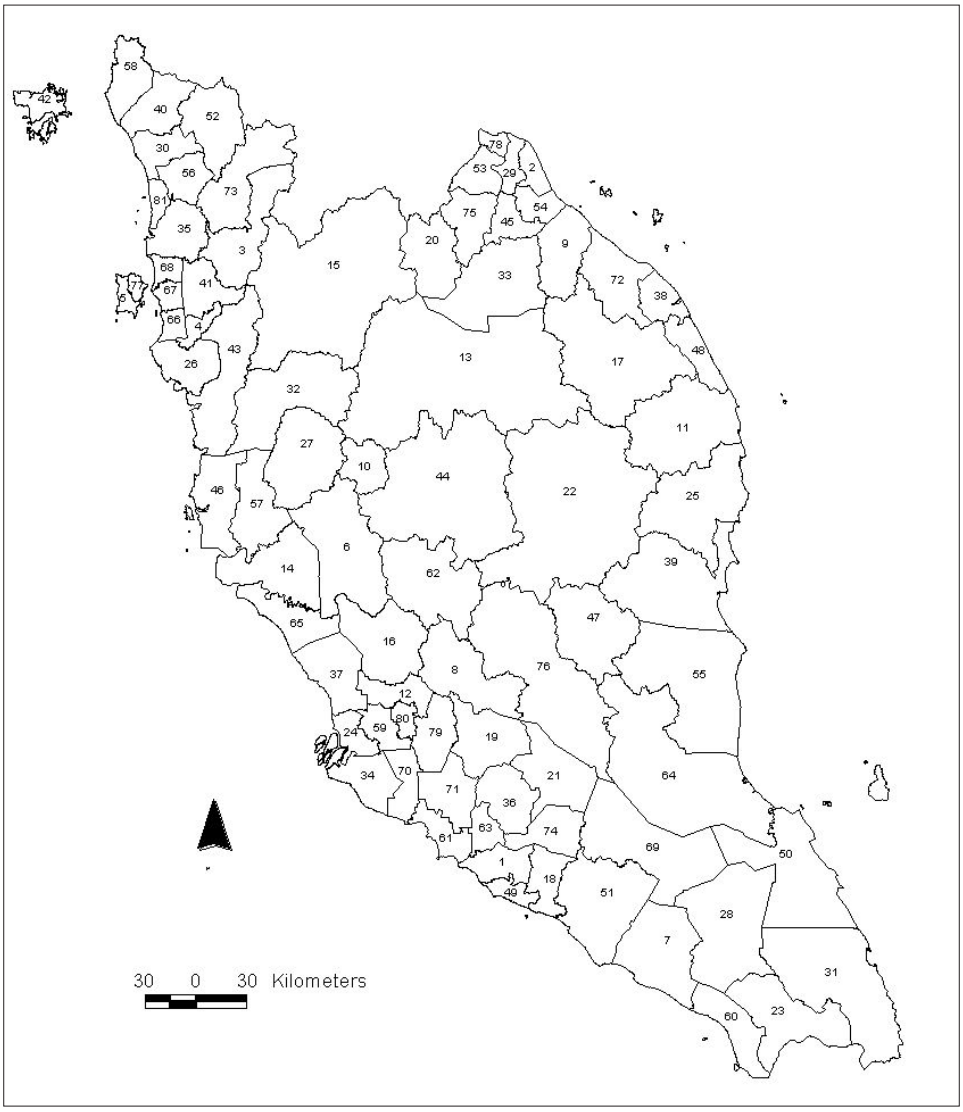

Figure 1 The study area – districts in Peninsular Malaysia

To examine the effects of data aggregation, the measure of population concentration was computed under two alternative conditions of analysis level: national (Peninsular Malaysia) and regional. At the national level, Hoover indices were calculated using three data aggregation levels: regional, state and district. Meanwhile for the regional level analysis, effects of data aggregation were examined using district and state level data. Following Otterstrom (2001), the districts were analysed in national, regional and states groupings to give a comparable set of relative values which can be interpreted according to the appropriate scale-sensitive perspective.

Table 1 Distribution population among states and districts in Peninsular Malaysia 1980–2000

State and administrative district	Population		
	1980	1991	2000
JOHOR	**1,580,423**	**2,069,740**	**2,565,701**
• Batu Pahat (7)	274,625	294,056	335,368
• Johor Baharu (23)	406,871	704,471	1,064,881
• Keluang (28)	179,791	224,424	254,631
• Kota Tinggi (31)	114,267	174,425	192,220
• Mersing (50)	42,208	63,643	67,557
• Muar (51)	291,129	301,804	328,695
• Pontian (60)	121,031	129,356	143,729
• Segamat (69)	150,501	177,561	178,620
KEDAH	**1,077,815**	**1,302,241**	**1,572,107**
• Baling (3)	104,858	114,485	124,947
• Bandar Baharu (4)	31,724	33,006	37,932
• Kota Setar (30)	279,567	322,354	354,431
• Kuala Muda (35)	192,308	254,372	339,898
• Kubang Pasu (40)	129,808	157,963	186,265
• Kulim (41)	92,525	128,356	191,160
• Langkawi (42)	28,340	42,938	69,597
• Padang Terap (52)	40,428	50,726	55,899
• Pendang (56)	75,861	83,092	89,790
• Sik (73)	43,366	54,466	59,691
• Yan (81)	59,030	60,483	62,497
KELANTAN	**859,270**	**1,181,315**	**1,289,199**
• Bachok (2)	73,953	98,557	109,786
• Gua Musang (13)	18,578	63,816	74,988
• Jeli (20)	23,352	32,720	36,057
• Kota Bharu (29)	275,986	366,770	400,321
• Kuala Krai (33)	62,301	90,830	91,619
• Machang (45)	58,040	71,584	77,921
• Pasir Mas (53)	118,153	150,035	162,296
• Pasir Putih (54)	80,959	96,348	104,734
• Tanah Merah (75)	61,996	94,611	101,450
• Tumpat (78)	85,952	116,044	130,027

cont.

Table 1 (*continuation*)

State and administrative district	Population		
	1980	1991	2000
MELAKA	**446,769**	**506,321**	**602,867**
• Alor Gajah (1)	113,083	116,653	131,870
• Jasin (18)	87,523	92,771	101,775
• Melaka Tengah (49)	246,163	296,897	369,222
NEGERI SEMBILAN	**551,442**	**692,897**	**830,080**
• Jelebu (19)	36,730	40,012	37,120
• Jempol (21)	67,159	122,033	125,151
• Kuala Pilah (36)	67,345	68,180	63,176
• Port Dickson (61)	83,561	92,171	106,919
• Rembau (63)	36,350	34,823	36,809
• Seremban (71)	202,790	263,383	383,982
• Tampin (74)	57,507	72,295	76,923
PAHANG	**768,801**	**1,045,003**	**1,231,176**
• Bentong (8)	72,865	83,965	97,467
• Cameron Highlands (10)	21,502	25,555	28,050
• Jerantut (22)	59,043	74,547	81,215
• Kuantan (39)	170,573	255,974	344,706
• Lipis (44)	56,996	68,276	73,391
• Maran (47)	91,187	110,264	112,626
• Pekan (55)	50,058	86,179	98,400
• Raub (62)	64,414	73,085	79,432
• Rompin (64)	38,975	80,251	101,877
• Temerloh & Bera (76)	143,188	186,907	214,012
PERAK	**1,743,655**	**1,877,471**	**2,030,382**
• Batang Padang (6)	136,473	154,686	152,137
• Hilir Perak (14)	203,028	202,059	191,098
• Hulu Perak (15)	71,372	81,636	82,195
• Kerian (26)	155,765	148,720	152,651
• Kinta (27)	564,886	627,899	751,825
• Kuala Kangsar (32)	146,292	146,684	154,048
• Larut & Matang (43)	249,550	271,882	273,321
• Manjung (46)	143,610	168,331	191,004
• Perak Tengah (57)	72,679	75,574	82,103

cont.

Table 1 (*continuation*)

State and administrative district	Population		
	1980	1991	2000
PERLIS	**144,782**	**183,824**	**198,335**
• Perlis (58)	144,782	183,824	198,335
PULAU PINANG	**900,772**	**1,064,166**	**1,225,501**
• Barat Daya (5)	76,390	122,764	159,019
• S.P.Selatan (Nibong Tebal) (66)	71,558	84,771	117,208
• S.P.Tengah (Bkt. Mertajam) (67)	161,975	236,270	291,876
• S.P.Utara (Butterworth) (68)	199,449	224,647	243,316
• Timur Laut (77)	391,400	395,714	414,082
SELANGOR	**1,426,250**	**2,297,159**	**3,947,527**
• Gombak (12)	166,059	352,649	553,410
• Kelang (24)	279,349	406,994	648,918
• Hulu Selangor (16)	81,679	82,814	142,771
• Kuala Langat (34)	101,578	130,090	189,983
• Kuala Selangor (37)	110,366	123,052	157,288
• Petaling (59)	360,056	633,165	1,181,034
• Sabak Bernam (65)	103,261	99,824	110,713
• Sepang (70)	46,025	54,671	97,896
• Ulu Langat (79)	177,877	413,900	865,514
TERENGGANU	**525,255**	**766,244**	**879,691**
• Besut * (9)	79,253	107,900	120,538
• Dungun (11)	58,360	102,897	128,582
• Hulu Terengganu (17)	43,459	56,986	62,262
• Kemaman (25)	64,899	111,901	136,502
• Kuala Terengganu * (38)	203,979	274,489	298,149
• Marang * (48)	45,641	69,637	83,165
• Setiu (72)	29,664	42,434	50,493
KUALA LUMPUR Federal Territory (80)	**916,610**	**1,145,342**	**1,297,526**
Peninsular Malaysia	**10,941,844**	**14,141,723**	**17,670,692**
MALAYSIA	**13,136,109**	**17,563,420**	**22,202,614**

Note: Administrative district boundary is based on 1991 Census. For Census 2000, district of Temerloh was divided into 2 new districts: Temerloh dan Bera.

Number in parenthesis corresponds to the district ID number as shown in Figure 1.0.

**Source*: Department of Statistics, Malaysia (2001).

The national level includes all districts in the Peninsular Malaysia. The regional level is defined by the breakdown of districts into one of the four regions following the National Regional Development Strategy (MPSP, 2001). For this purpose Peninsular Malaysia is divided into four main development regions: Northern (Perlis, Kedah, Penang and Perak), central (Selangor and Federal Territory Kuala Lumpur), southern (Negeri Sembilan, Melaka and Johor) and eastern (Kelantan, Terengganu and Pahang) (Figure 2). Therefore the regional level analysis of this research could be useful for monitoring the impact of regional development on population distribution. Table 2 shows the distribution of population among the regions. The state level is defined by breakdown of districts according to the official administrative boundary of each state.

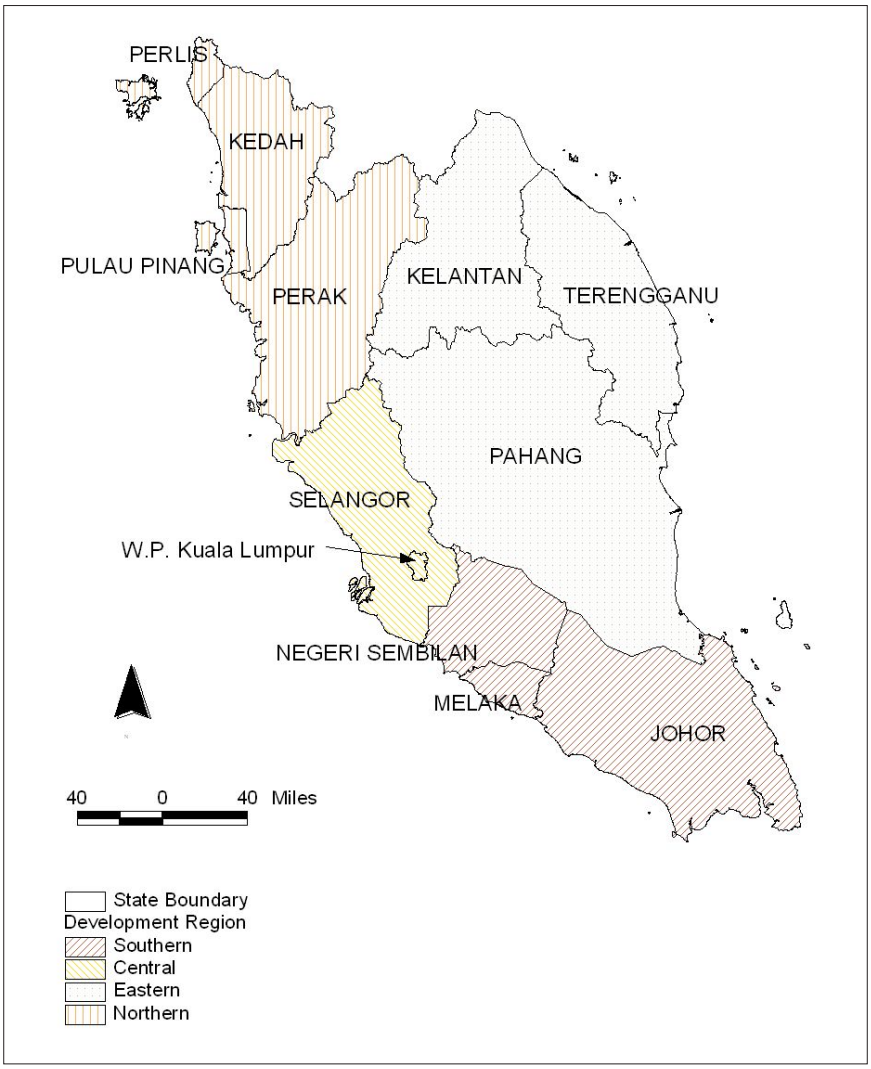

Figure 2 Development region of Peninsular Malaysia

Table 2 Distribution of total population by development region Peninsular Malaysia 1980-2000

Region	Population*		
	1980	1991	2000
Northern	3,867,024	4,427,702	5,026,325
Central	2,345,860	3,442,501	5,245,053
Southern	2,578,634	3,268,958	3,998,648
Eastern	2,153,316	2,992,562	3,400,066
Pen. Malaysia	10,944,844	14,131,723	17,670,092

* *Source*: Department of Statistics, Malaysia (2001).

The analysis was carried out using ArcView geographical information system (GIS) software.

3.0 RESULTS

Table 3 and Figure 3 show the results of the calculation of Hoover population concentration index at the national level for the three levels of data aggregation. Using the district level data, the Hoover indices for Peninsular Malaysia were 49.24, 48.85 and 51.57 for the year 1980, 1991 and 2000 respectively (Table 3). This indicates that for a period of the 20 years, the population of Peninsular Malaysia as a whole was relatively uniform, not too concentrated nor dispersed. However, there was a slight 'break' between 1980 and 1991 whereby the Hoover index slightly decreased from 49.24 to 48.85 indicating to a slight deconcentration in population in that decade. However, once the data were aggregated to regional and state level, the Hoover index becomes smaller (about half) indicating the population distribution is more dispersed.

Table 3 Distribution of population concentration index at the national level for different data aggregation levels

Aggregation level	Population concentration index		
	1980	1991	2000
District data	49.24	48.85	51.57
State data	28.71	29.70	33.81
Regional data	28.71	27.21	29.14

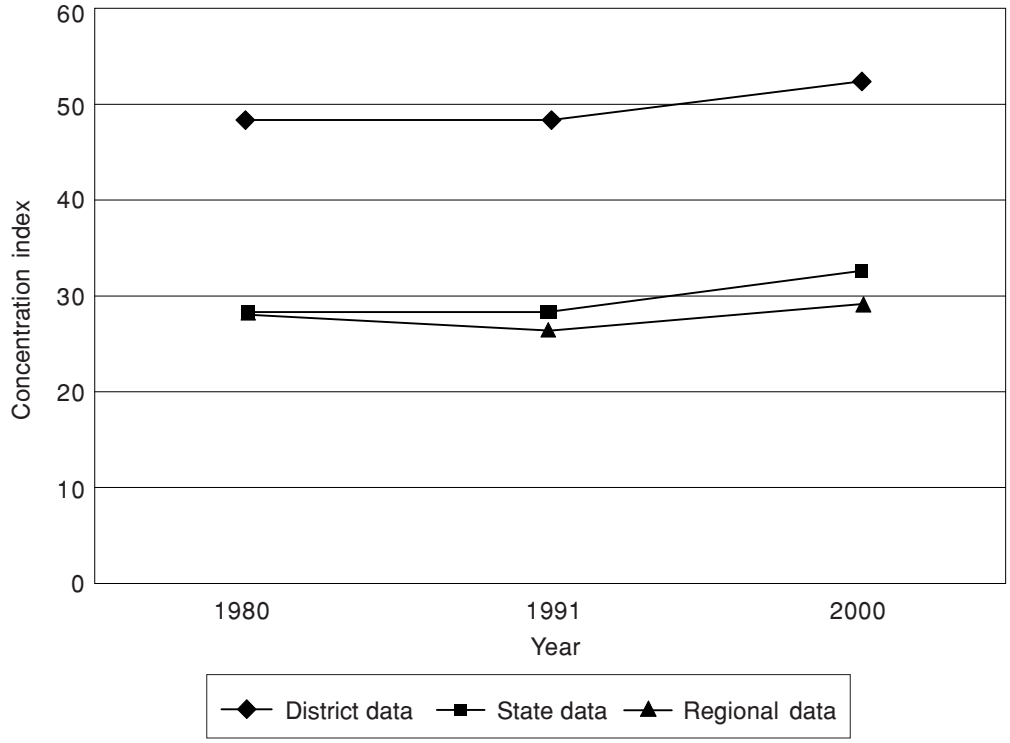

Figure 3 Distribution of population concentration index for the national level at different data aggregation

Interestingly, eventhough regions are larger than the state, the indices were relatively similar. This indicates that the sensitivity of the method levels off at certain data aggregation level.

However, the trends of population concentration showed by the regional and state levels data were different. Using the regional population data, the population concentration indices decreased slightly in 1991 but increased slightly again in 2000. On the other hand, analysis using state level data indicates that the population concentration of Peninsular Malaysia continuously increases within the 20-year period. Thus, the results of this study clearly indicates that data aggregation levels affect the measure of population concentration and are likely to produce totally different trend. In terms of population, what this result indicates is that there are intraregional migrations in that particular region, where people migrated from one state to another within the same region.

To further enhance the results, analysis at the regional level was carried out. As described earlier, Peninsular Malaysia is divided into four development regions, each with varying sizes and numbers of states and districts. Table 4 and Figure 4 show the population concentration indices between and within regions for the two data aggregation levels, namely the states and districts. Similar to the analysis at the national

Table 4 Distribution of population concentration index for each development region 1980 – 2000

Region	Population concentration index		
	1980	1991	2000
North			
District data	39.58	41.09	43.38
State data	21.38	22.72	24.72
Central			
District data	46.18	48.28	49.21
State data	36.37	30.34	21.91
Eastern			
District data	46.44	43.98	43.64
State data	20.61	21.39	20.11
Southern			
District data	33.75	33.01	37.47
State data	11.28	9.44	9.03

level, the effects of data aggregation on the calculation of population concentration index at the regional level also vary. For the northern region, both concentration indices show increasing concentration over the year eventhough the state data indicates a much lower concentration. However for the other three regions, the trends of population concentration using the district data are different than the state data. For the central region, the district data indicates increasing concentration while the state data, on the other hand, indicates decreasing concentration. This means that there is a dispersal of population between states, but increasing concentration within regions. For the eastern region, which comprised of Kelantan, Terengganu and Pahang, the district data indicates a decreasing population concentration over the 20 years period. The state data, on the other hand, indicates a slight increase in population concentration between 1980 and 1991, then decreases again in 2000. For the southern region, the district data generally indicates an increasing trend of population concentration. The state data,

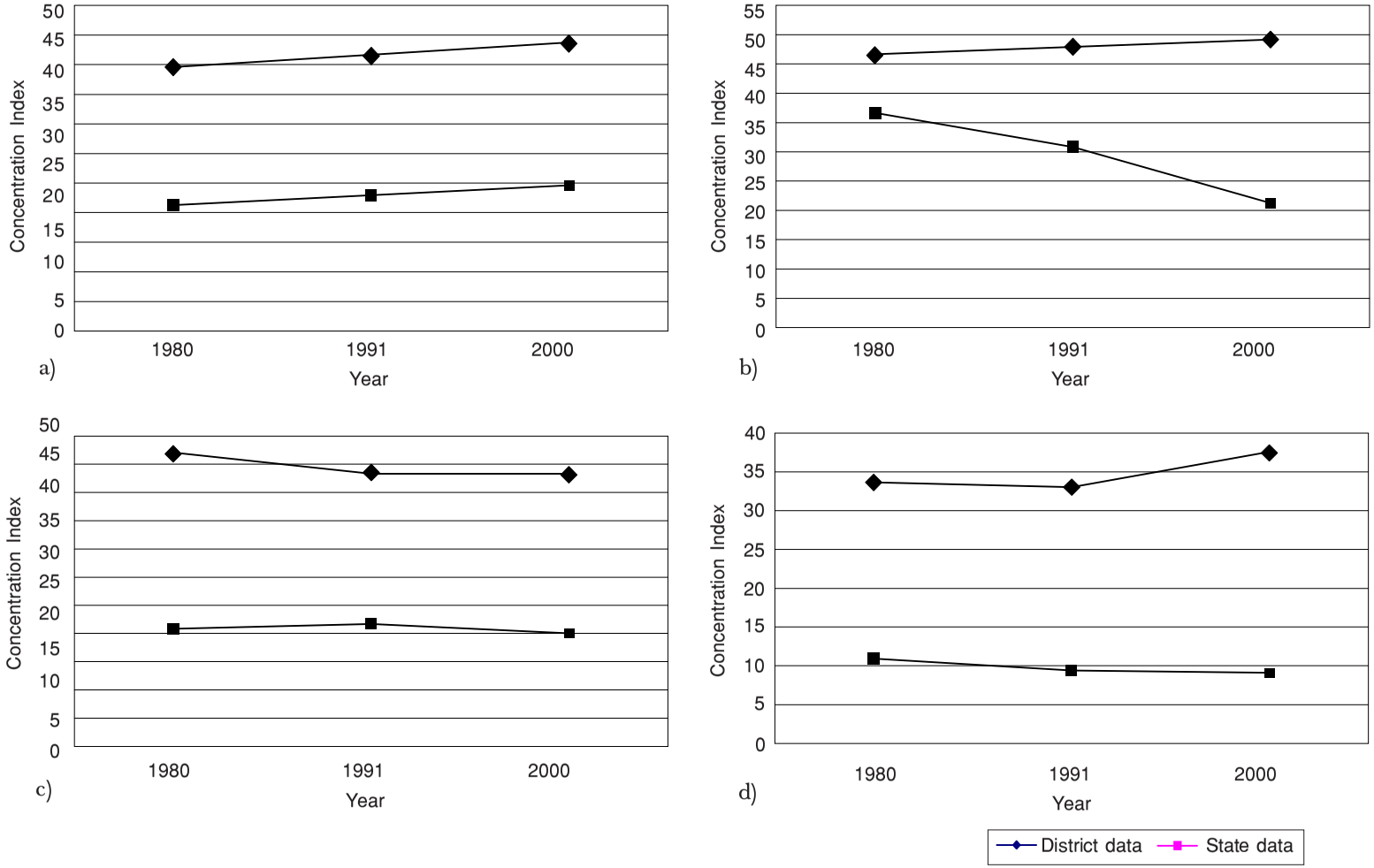

Figure 4 Distribution of population concentration index for the regional level at different data aggregation a). northern b). central c). eastern and d). southern

however indicates a decreasing trend in population concentration. This means that there is a dispersal of population between states, but movement of the population are concentrated only to selected districts.

In summary, this study has found that data aggregation will affect the calculation of population concentration index, usually towards lower concentration. However, the trends of population concentration produced by aggregated data are not necessarily similar to the more detailed data.

4.0 DISCUSSION

This study has successfully shown that data aggregation could affect the measurement of population concentration. This phenomenon is known as the Modifiable Areal Unit Problem (MAUP), an issue that has been previously raised by a number of authors like Openshaw (1983). The results of this study confirmed the word of caution issued by Duncan (1957) pertaining to the use of index of concentration. In general, the value of the index is directly related to the number of areal units into which the territory is sub-divided, or inversely related to the average size of the units. An index computed for a given set of areal units cannot be larger than the index computed for a set which comprises sub-divisions of the first set. Thus, the index based on districts must be at least as great as the index based on regions or states (combinations of districts), and will be greater if there is any unevenness of distribution by districts within sub-regions. As a result, the index provides no unique answer to the question of what degree of population concentration characterizes a territory (such as nation). For this reason, any index value must be interpreted relative to the system of sub-areas on which it is based. Furthermore, the index does not give a unique answer to the question of whether the unevenness of distribution is increasing or decreasing. This is evident from the series shown in Table 3 and 4 that changes in contrary directions. Nonetheless, this does not mean that the measure of population concentration is empirically meaningless. Looking closely at the results, such measure is useful for comparative study especially of temporal changes.

As described in the preceding section, the decreasing indices based on region reflects the spread of population over the nation. In other words, there has been migration between regions over the periods. The decreasing indices based on districts between 1980 and 1991 may reflect local deconcentration of population within metropolitan (urban) areas, and the increasing indices indicate urban and metropolitan concentration of population. The contrary results obtained with alternative indices (for example central region in Table 4) may then reflect a basic ambiguity inherent in any concept of concentration that does not specify the system of areal units to which it refers, rather than a defect in the operational definition of the measure of concentration (Duncan, 1957). This study used Hoover index, a measure that had been used quite extensively in the study of population distribution. However, there are many other

methods that could be used for similar purpose such as the Gini index, Lorenz curve and entropy index. To the knowledge of the author, no such study has been carried out so far, and could be a subject for further research.

5.0 SUMMARY AND CONCLUSION

This article has described a study on the effects of data aggregation on the measure of spatial concentration of population. This study has found that aggregated data will produce a much lower measure of population concentration. However, the trends of population concentration shown by the detailed data might not necessarily be replicated by the aggregated data. Depending upon the pattern of population distribution, the trend produced by the aggregated data could be totally in different direction than those produced by the more detailed data. The resulting patterns will greatly depend on the movement of population on the smaller spatial scale especially those within the same state or region. Therefore, analysis of population concentration using different data aggregation could provide a useful early indication on the movement of population between and within region or state prior to a detailed study on migration pattern.

Population concentration is more observable using detailed data as compared to aggregated data. In addition, this study also found that data aggregation will only affect the measure of population concentration up to certain spatial scale, beyond which the effects become minimal. Therefore, it is not sufficient to look at population distribution only at one geographical level (usually at the national level). With the advancement of technology in handling spatial data (geographical information system for example) and detailed digital data are easily available, detailed analysis and variation of population spatial distribution are becoming easier to carry out in the future.

ACKNOWLEDGEMENTS

The authors would like to thank the reviewers for their helpful comments. Any remaining errors can only be the authors'.

REFERENCES

Department of Statistics, Malaysia. 2001. *Census 2000 Preliminary Count.*

Duncan, O. D., 1957. "The Measurement of Population Distribution". *Population Studies*. 11(1): 27-45.

Duncan, O. D., R. P.Cuzzort, and B. Duncan. 1961. *Statistical Geography*. Glencoe, Illinois: The Free Press.

Hoover, E. 1941. "Interstate redistrubution of population, 1850-1940". *Journal of Economic History*. 1: 199-205.

Lichter, Daniel T. 1985. "Racial Concentration and Segregation Across US Counties, 1950 - 1980". *Demography*. 22(4): 603-609.

Majlis Perbandaran Pulau Pinang (MPPP). 2001. *MPPP Structure Plan (Modification) Draft Report*. Majlis Perbandaran Pulau Pinang.

Majlis Perbandaran Seberang Perai (MPSP). 2001. *MPSP Structure Plan (Modification) Draft Report*. Majlis Perbandaran Seberang Perai.

Openshaw, S. 1983. The modifiable areal unit problem. *Concepts and techniques in modern geography*. 38.

Otterstrom, Samuel M. 2001. "Trends in national and regional population concentration in the United States from 1790 to 1990: from the frontier to the urban transformation". *The Social Science Journal*. 38: 393-407.

Portnov, Boris A., and Pearlmutter, David. 1999. "Sustainable urban growth in peripheral areas". *Progress in Planning*. 52: 239-308.

Souza, M.L.de. 2001. "Metropolitan deconcentration, socio-political fragmentation and extended suburbanisation: Brazilian urbanization in the 1980s and 1990s". *Geoforum*. 32: 437-447.

Vining, D.R., and Strauss, A. 1977. "A demonstration that the current deconcentration of population in the United States is a clean break with the past". *Environment and Planning A*. 9: 751-758.