PERLOMBONGAN DATA JUJUKAN

STQD6414 PERLOMBONGAN DATA

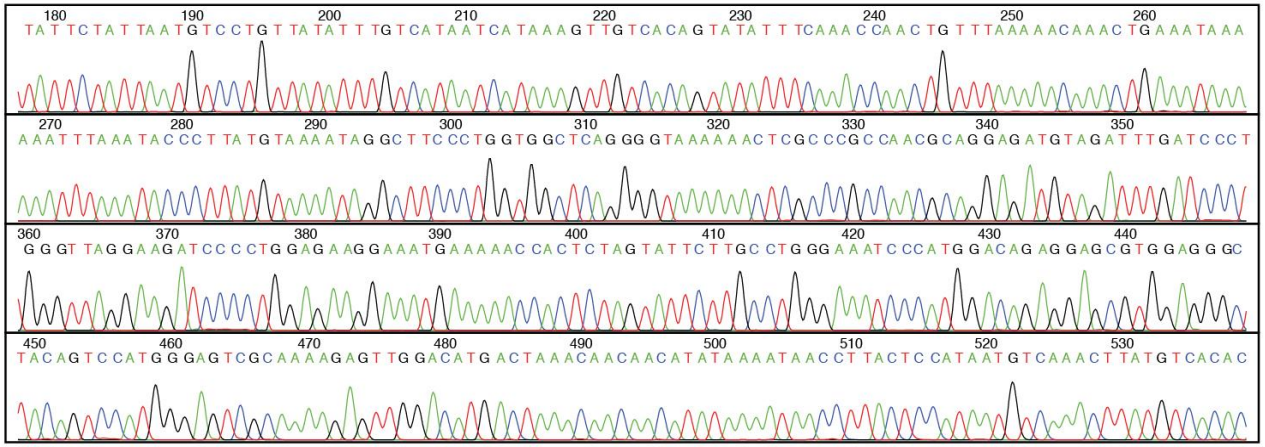
Prof. Madya Dr. Nurulkamal Masseran

Jabatan Sains Matematik

Universiti Kebangsaan Malaysia

PENGENALAN:

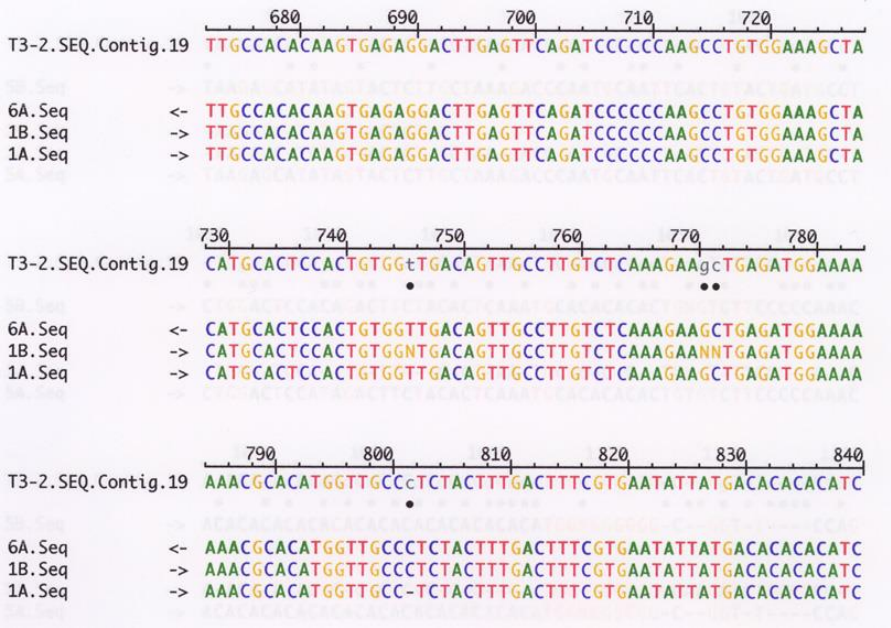
- Topik ini akan membincangkan berkaitan analisis data jujukan berkategori.
- Dalam data jujukan, posisi untuk setiap keadaan (*state*) secara berturut-turut memberikan tafsiran dalam sebutan umur, tarikh, masa yang berlalu (*elapsed time*) ataupun jarak dari permulaan jujukan.
- Umumnya, data jenis ini merujuk kepada cerapan-cerapan individu atau entiti tertentu dalam suatu tempoh masa.
- Objektif utama adalah untuk menganalisis tingkah laku bagi jujukan keadaan bagi entiti/individu yang dicerap.

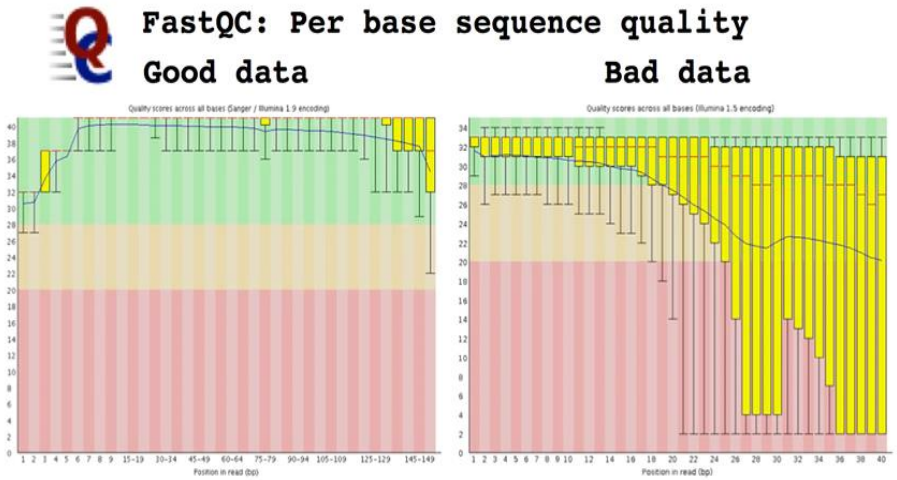


DNA sequence data from an automated sequencing machine

INTRODUCTION:

- Dalam topik ini, perbincangan kita akan memfokus kepada analisis jujukan bagi data trajektori hayat (*life trajectory*).
- Namun, kebanyakan konsep dan teknik bagi analisis jujukan boleh juga digunakan dalam pelbagai domain bidang, antaranya; biologi, kualiti kontrol, data text, data log-web, dan lain-lain.





DATA JUJUKAN:

- Jujukan merupakan objek yang kompleks dan ianya memerlukan teknik khas untuk menganalisis data jenis ini.

- Antara persoalan yang menarik untuk dianalisis dalam data jenis jujukan:

i) Apakah ciri-ciri yang wujud dalam data jujukan?

ii) Apakah penunjuk-penunjuk (*indicators*) yang boleh digunakan untuk ukuran data jujukan?

iii) Apakah plot-plot yang sesuai untuk mengvisualkan data jujukan?

iv) Bagaimana untuk membandingkan ciri kesamaan antara beberapa data jujukan?

JUJUKAN KEADAAN:

- Jujukan keadaan merupakan konsep penting yang digunakan untuk menganalisis trajektori hayat.
- Contoh: sejarah pekerjaan, sejarah tahap pesakit, perjalanan hidup bersekedudukan (*cohabitation life*) dan lain-lain.

- Berdasarkan data jujukan, maklumat yang boleh diperolehi:

i) Ciri-ciri norma social atau trajektori piawai dalam perjalanan hidup.

ii) Tingkah laku yang menyimpang daripada trajektori piawai.

iii) Corak evolusi perjalanan hidup dari semasa ke semasa.

iv) Ciri-ciri trajektori kehidupan terhadap faktor berkaitan jantina, budaya asal sosial, dan lain-lain.

JUJUKAN KEADAAN:

- Analisis jujukan keadaan bertujuan meringkaskan dan mengkategorikan pola jujukan kepada beberapa kumpulan tertentu yang mempunyai sifat yang serupa.

- Antara teknik penting dalam analisis data jujukan:

i) Penggunaan penunjuk ringkasan statistik.

ii) Pengvisualan data.

iii) Pengkelompokan data.

iv) Perbandingan ciri jujukan.

- Hasil yang diperolehi akan memberikan maklumat untuk analisis lanjutan yang melibatkan pelbagai kaedah statistik lain yang lebih kompleks.

JENIS-JENIS DATA JUJUKAN:

- Terdapat beberapa jenis data jujukan:

i) Format Jujukan-Keadaan (STS).

ii) Format Jujukan-Keadaan-Kekal (SPS).

iii) Format Peristiwa-Bertanda-Masa (TSE).

iv) Format SPELL.

v) Dan lain-lain.

Code	Example										
STS	**Id**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**
	101	S	S	S	M	M	MC	MC	MC	MC	D
	102	S	S	S	MC	MC	MC	MC	MC	MC	MC
SPS (1)	**Id**	**State 1**		**State 2**		**State 3**		**State 4**		**State 5**	
	101	(S,3)		(M,2)		(MC,4)		(D,1)			
	102	(S,3)		(MC,7)							
SPS (2)	**Id**	**State 1**		**State 2**		**State 3**		**State 4**		**State 5**	
	101	S/3		M/2		MC/4		D/1			
	102	S/3		MC/7							
DSS	**Id**	**State 1**		**State 2**		**State 3**		**State 4**		**State 5**	
	101	S		M		MC		D			
	102	S		MC							
TSE	**id**	**time**	**event**								
	101	21	Marriage								
	101	23	Child								
	101	27	Divorce								
	102	21	Marriage								
	102	21	Child								
SPELL	**id**	**index**	**from**	**to**	**status**						
	101	1	18	20	Single						
	101	2	21	22	Married						
	101	3	23	26	Married w Children						
	101	4	27	..	Divorced						
	102	1	18	20	Single						
	102	2	21	27	Married w Children						

PENUNJUK RINGKASAN STATISTIK:

- Antara penunjuk ringkasan statistik yang penting dalam analisis data jujukan:

i) Min (purata) masa proses berada dalam setiap keadaan.

ii) Min (purata) masa proses berada dalam setiap keadaan bagi kumpulan tertentu.

iii) Bilangan transisi (peralihan).

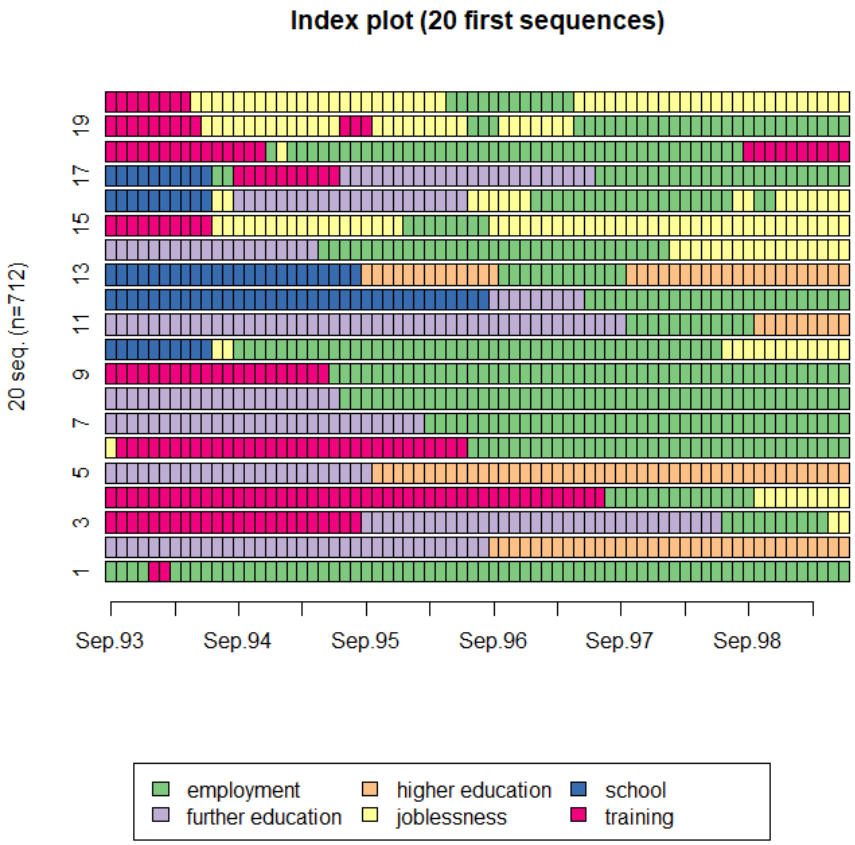
iv) Kadar peralihan.

v) Keadaan peralihan yang bergantung terhadap masa.

vi) Dan lain-lain.

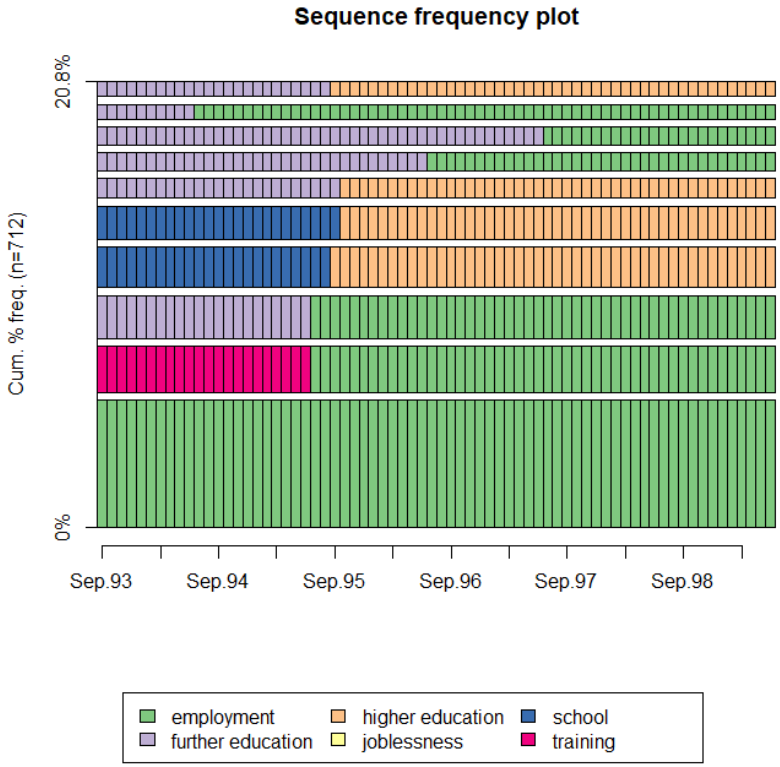
PLOT INDEKS JUJUKAN:

- Plot indeks jujukan boleh digunakan untuk menggambarkan tingkahlaku jujukan keadaan dalam data.
- Plot ini diwakili oleh kotak bertindan mendatar yang diwarnakan berdasarkan keadaan yang berbeza.
- Lebar bar mendatar mewakili perkadaran setiap kekerapan.
- Setiap bar dengan warna dan panjang yang berbeza memaparkan maklumat tentang perubahan longitudinal individu dari satu keadaan kepada keadaan lain.



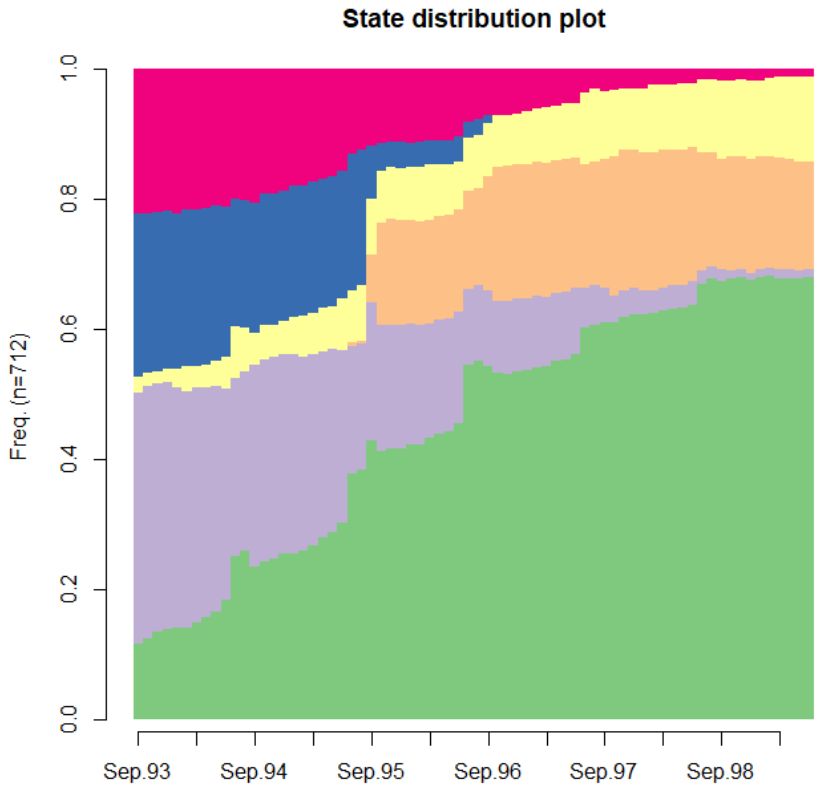
PLOT JUJUKAN KEKERAPAN:

- Kekerapan jujukan merujuk kepada bilangan dan peratus kekerapan jujukan yang disusun mengikut susunan kekerapan yang menurun.
- Plot kekerapan jujukan memberikan paparan grafik bagi kekerapan jujukan dengan lebar bar adalah berkadar dengan kekerapannya.



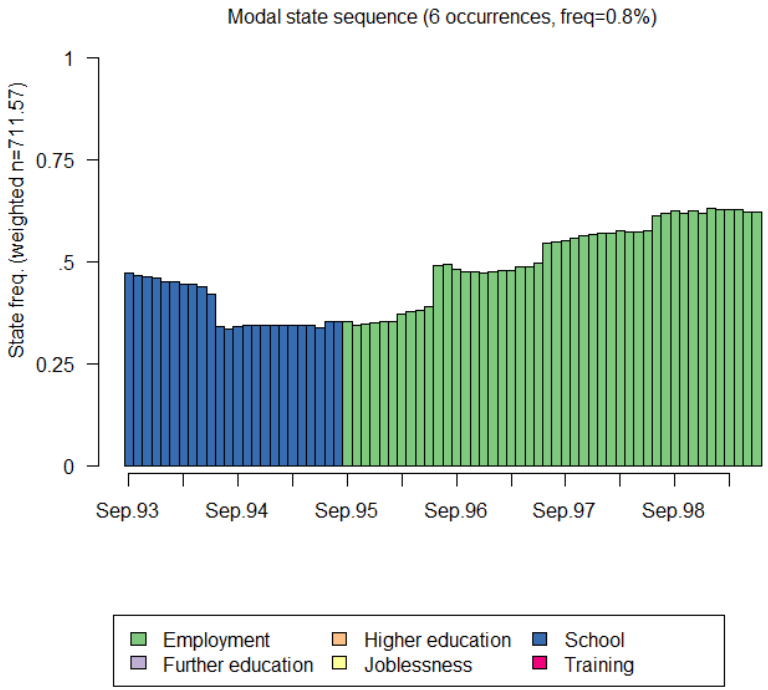
PLOT TABURAN KEADAAN:

- Plot ini memaparkan corak umum keseluruhan set trajektori dalam data jujukan.
- Ia memberikan paparan agregat untuk ciri rentas lintang (*transversal characteristics*) bagi data jujukan.



PLOT KEADAAN MODAL:

- Plot ini memberikan maklumat tentang jujukan bagi keadaan yang paling kerap berlaku pada setiap kedudukan.
- Ia juga memaparkan maklumat bilangan keberlakuan keadaan modal dalam data jujukan.



CIRI JUJUKAN: INDEKS ENTROPI

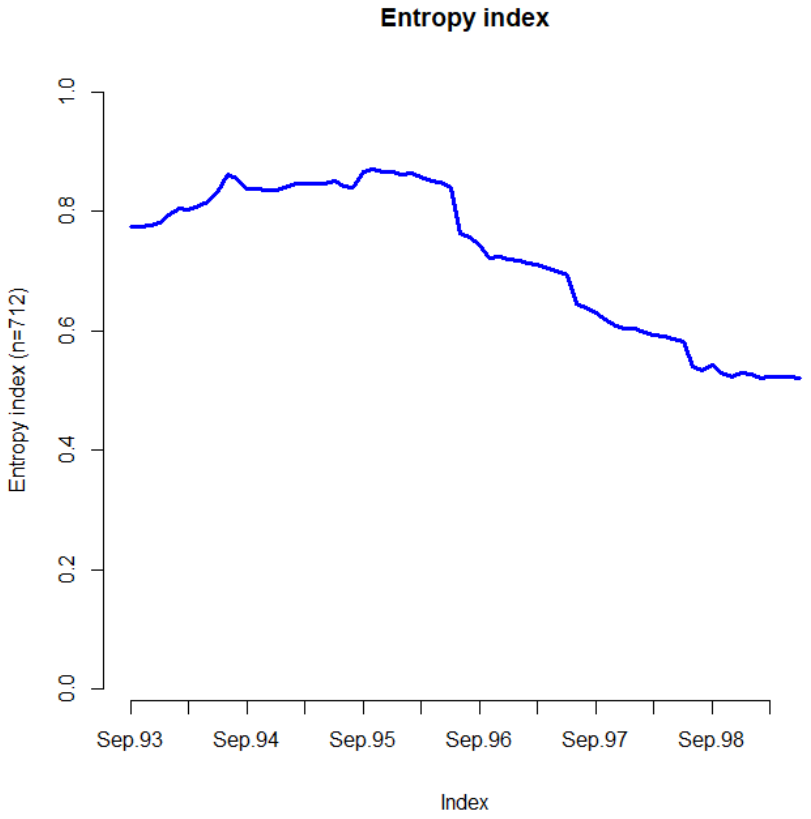
- Entropi merupakan ukuran terhadap variasi keadaan dalam data jujukan
- Indeks Entropi bagi data jujukan boleh diperolehi menerusi:

$$h(p_1, \ldots, p_a) = -\sum_{i=1}^{a} p_i \log(p_i)$$

- dengan p_i ialah perkadaran entiti dalam keadaan-*i*, *a* ialah saiz data jujukan.
- Jika nilai entropi=0, menunjukkan bahawa semua entiti berada dalam keadaan yang sama (variasi adalah 0).
- Jika nilai entropi tinggi, menunjukkan bahawa perkadaran entiti dalam setiap keadaaan adalah hampir sama (variasi tinggi).

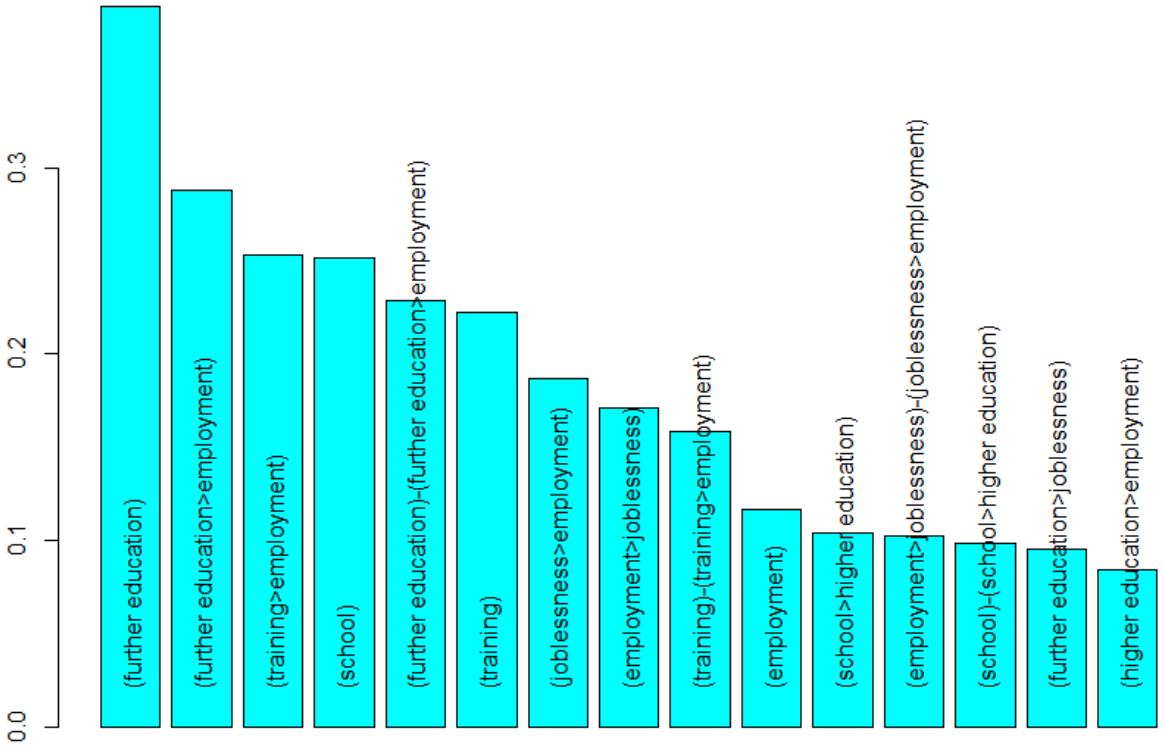
ENTROPI RENTAS LINTANG:

- Plot entropi rentas lintang memaparkan maklumat perubahan variasi keadaan-keadaan dalam data jujukan terhadap faktor masa.



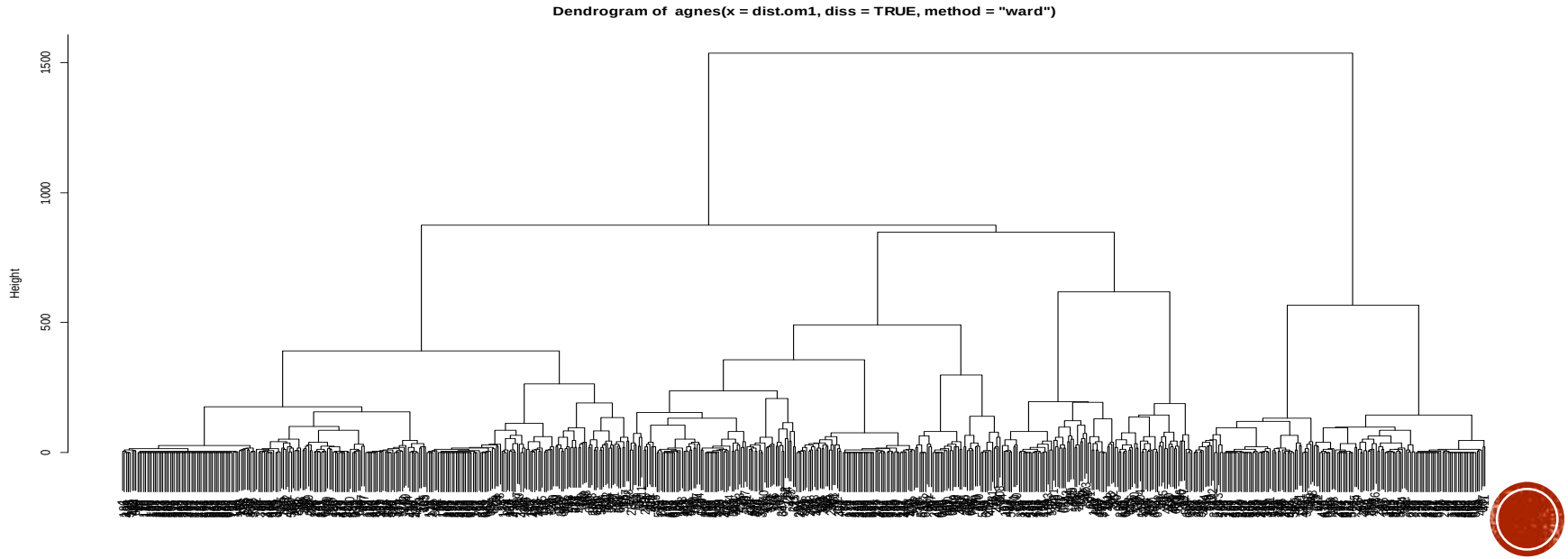
ANALISIS JUJUKAN PERISTIWA:

- Berbandinh data jujukan keadaan, kita boleh melihat data jujukan peralihan atau peristiwa.



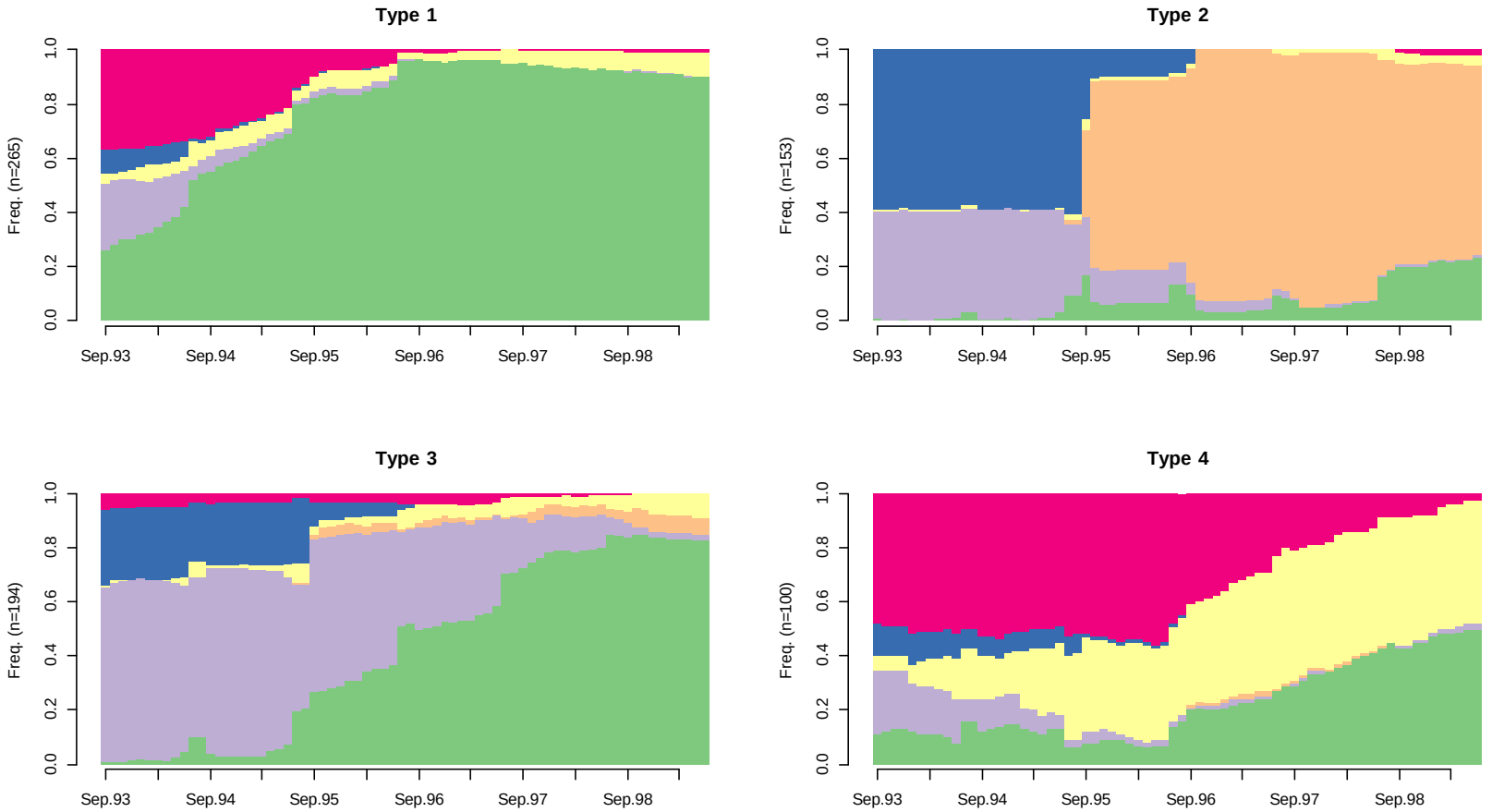
MENGKATEGORIKAN CORAK:

- Pengkelompokan corak memberikan maklumat tentang tipologi jujukan.
- Ianya dilakukan dengan mengukur kesamaan antara jarak berpasangan antara jujukan
- Teknik ini adalah berdasarkan algoritma padanan optimum (*optimal matching*).
- Setiap kelompok entiti kumpulan menunjukkan ciri trajektori yang serupa.

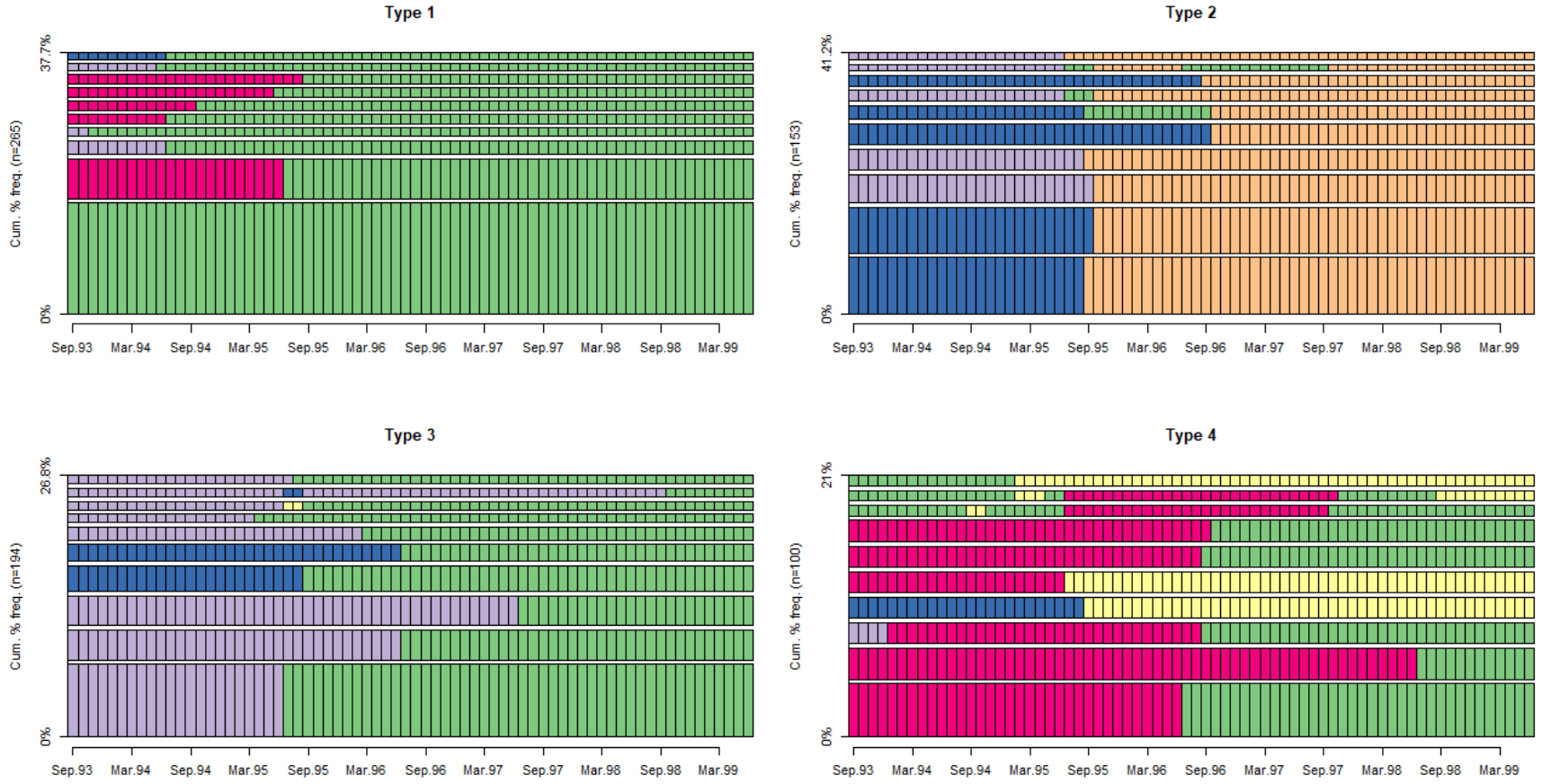


MENGKATEGORIKAN CORAK:

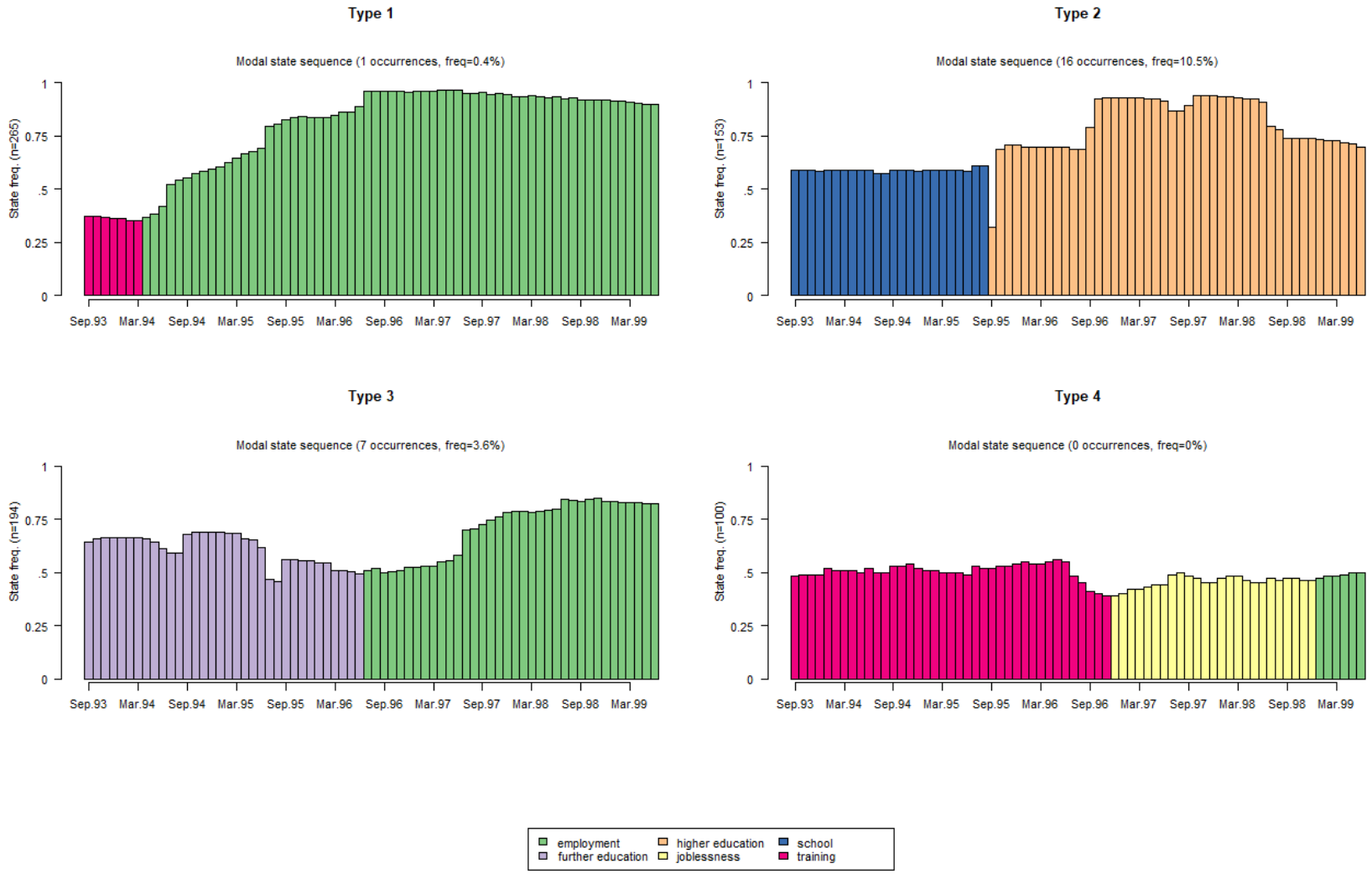
TABURAN KEADAAN



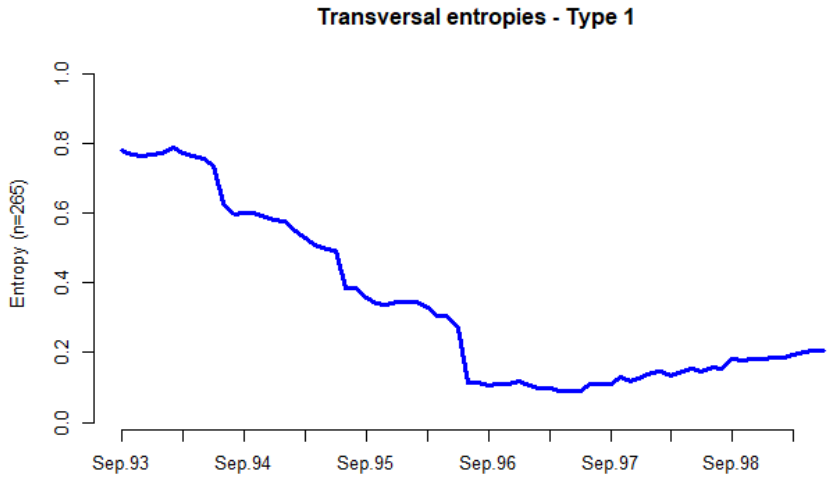
MENGKATEGORIKAN CORAK:
KEKERAPAN JUJUKAN

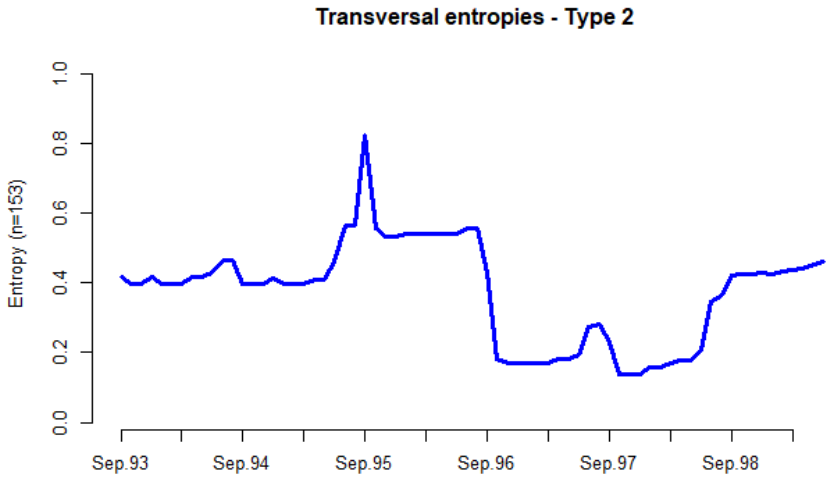


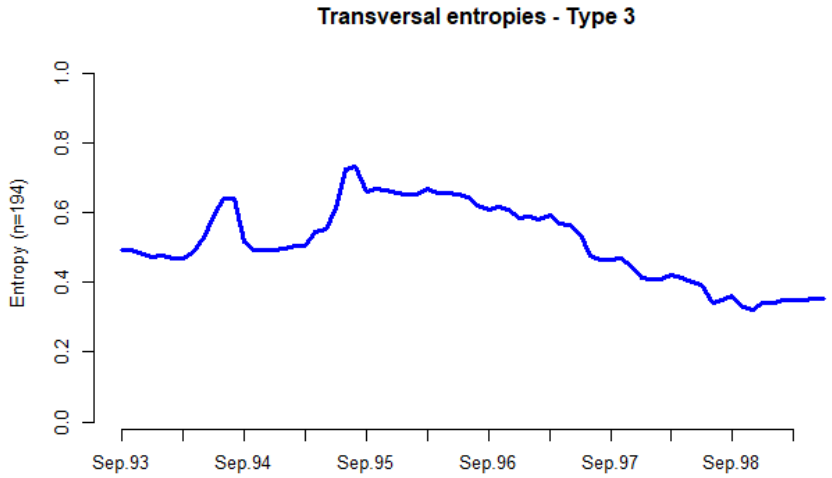
MENGKATEGORIKAN CORAK:
KEADAAN MODAL

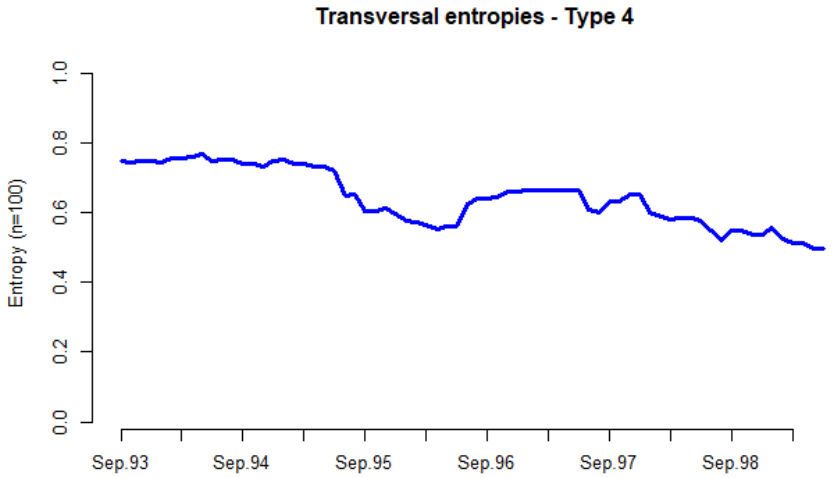


MENGKATEGORIKAN CORAK: ENTROPI RENTAS LINTANG



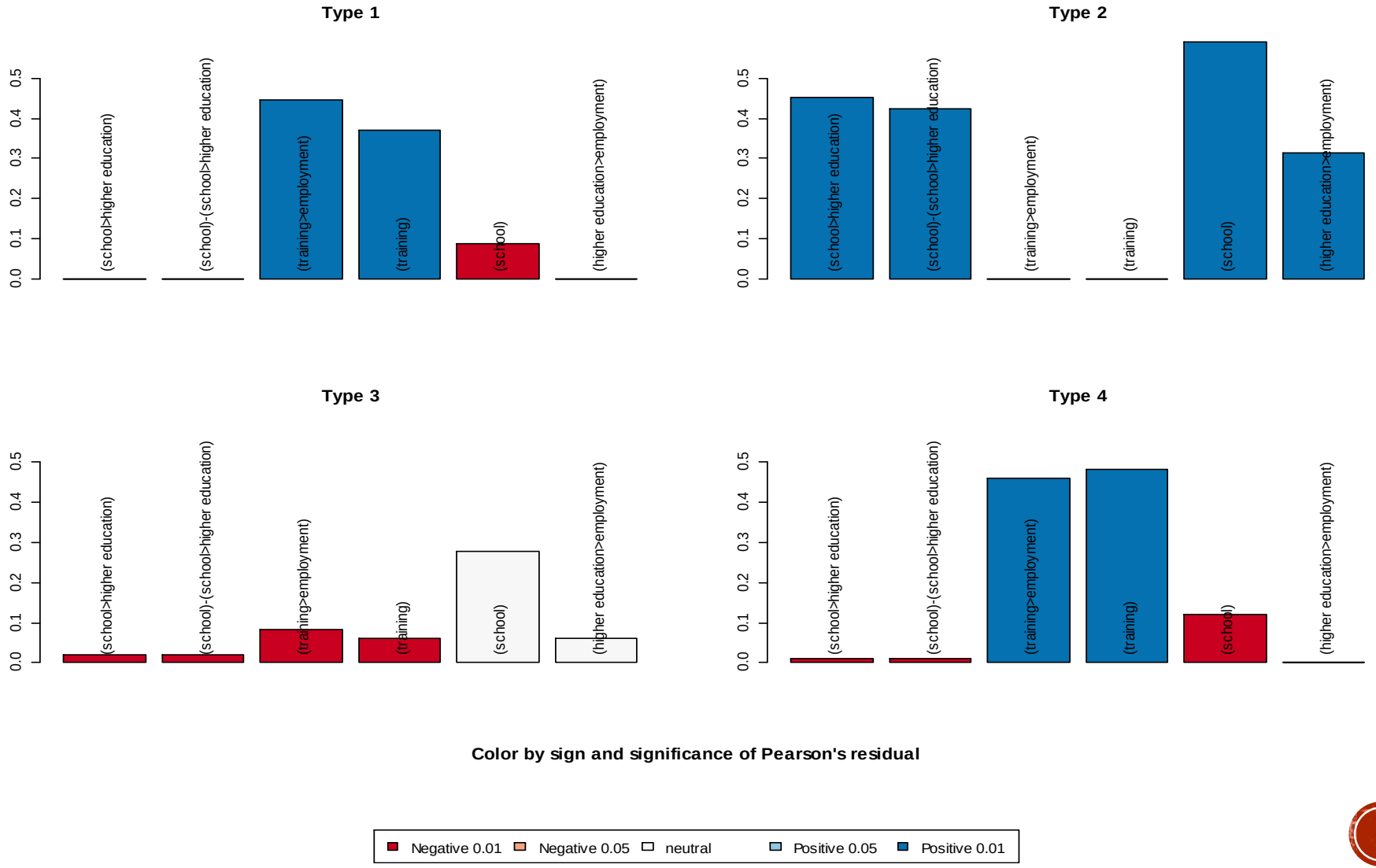






MENGKATEGORIKAN CORAK:

PEMBEZA TRANSISI



ANALYSIS JUJUKAN: PENDEKATAN LAIN

- Terdapat banyak pendekatan laian yang boleh digunakan untuk menganalisis data jujukan keadaan.

- Antaranya:

i) Analisis koresponden.

ii) Permodelan Markov.

iii) Analisis kemandirian.

iv) Analisis longitudinal.

v) Analisis panel data diskret.

vi) Dan lain-lain.

RUJUKAN:

- Curry, E. (2021). *Introduction to Bioinformatics with R: A Practical Guide for Biologists*. Boca Raton, Taylor & Francis.
- Gabadinho, A., Ritschard, G., Müller, N.S., Studer, M. (2011). Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software*, 40(4), 1–37.
- Gabadinho, A., Ritschard, G. (2016). Analyzing State Sequences with Probabilistic Suffix Trees: The PST R Package. *Journal of Statistical Software*, 72(3), 1–39.
- Melnykov, V. (2016). ClickClust: An R Package for Model-Based Clustering of Categorical Sequences. *Journal of Statistical Software*, 74(9), 1–34.
- Raab, M., Struffolino, E. (2022). *Sequence Analysis*. SAGE Publications.

TOPIK SETERUSNYA:

Perlombongan Data Reruang