使用機器學習進行術數量化實證研究設計

林聖軒

國立陽明交通大學統計學研究所副教授
國立陽明交通大學數據科學與工程研究所 合聘副教授

謝宜真

國立陽明交通大學統計學研究所碩士生

劉政猷

輔仁大學宗教研究所博士候選人

鄭志明

輔仁大學宗教研究所教授

摘要

中華傳統術數經過前人智慧長時間的累積，發展出非常精緻的理論體系與推命實務經驗。然而在大數據的時代，使用資料庫來檢驗這些術數方法在真實數據的表現是當代的趨勢。臺灣地區社會變遷基本調查(簡稱變遷調查)是台灣最大的社會調查資料庫，問卷內容包含了個人婚姻狀態、學歷、收入、家庭與親友鄰居的相處情形，這些內容都是術數界推命常見項目。在 1999 年以及 2004 年兩次變遷調查宗教組的問卷中有詢問受試者的生辰八字的題目，因此讓以生辰八字為主的推命術(包含紫微斗數以及子平八字等)有了得以經過數據量化的絕佳機會。在此我們提出一組研究設計的構想：使用變遷調查資料來進行紫微斗數論命法中推論婚姻狀態的量化實證研究。我們首先計畫請紫微斗數專家將古書中的各種對於婚姻推命數量化成可操作定義的「婚姻參數」，並將算法流程標準化為「婚姻傾向分數」；接著使用統計方法分析「婚姻參數」、「婚姻傾向分數」與真實婚姻狀態的相關性，以及使用機器學習模型建構「婚姻參數」與婚姻的預測模型，並計算其預測能力。本文提出的量化研究設計便能夠給予以生辰八字為主的推命數在進行量化研究時的重要的指引，期待能夠拋磚引玉，令各方術數大師未來願意一同合作得到最完整的推命術，更希望透過這樣的研究設計，讓各種術數成為資料科學上可重複驗證的命題。並且融合不同的學門及分析方式，截長補短，提升推命數在現代的適用性，讓傳統術數在學術界發揚光大。

關鍵詞：紫微斗數、婚姻、社會變遷調查、機器學習、效度研究、統計科學。

壹、前言

中華傳統術數經過前人智慧長時間的累積，發展出非常精緻的理論體系與推命實務經驗。然而在大數據的時代，使用資料庫來檢驗這些術數方法在真實數據的表現是當代的趨勢。臺灣地區社會變遷基本調查(簡稱變遷調查)是台灣最大的社會調查資料庫，問卷內容包含了個人婚姻狀態、學歷、收入、家庭與親友鄰居的相處情形，這些內容都是術數界推命常見項目。在 1999 年以及 2004 年兩次變遷調查宗教組的問卷中有詢問受試者的生辰八字的題目，因此讓以生辰八字為主的推命術(包含紫微斗數以及子平八字等)有了得以經過數據量化的絕佳機會。在此我們提出一組研究設計的構想：使用變遷調查資料來進行紫微斗數論命法中推論婚姻狀態的量化實證研究。我們首先計畫請紫微斗數專家將古書中的各種對於婚姻推命數量化成可操作定義的「婚姻參數」，並將算法流程標準化為「婚姻傾向分數」；接著使用統計方法分析「婚姻參數」、「婚姻傾向分數」與真實婚姻狀態的相關性，以及使用機器學習模型建構「婚姻參數」與婚姻的預測模型，並計算其預測能力。

貳、資料介紹與研究設計

一、資料庫描述

本研究資料取自變遷調查資料庫中第三期及第四期的第五次宗教組的問卷資料(簡稱「變遷調查宗教組」)。其資料為中央研究院民族學研究所依據羅啟宏先生所著之「台灣省鄉鎮發展類型之研究」的分層原則，自全台灣地區以分層抽樣的方式抽出訪問對象，再以電話訪問的方式由訪員向受訪者訪問有關家庭、教育、社會、政治、文化、傳播、宗教、人際關係及社會網絡等多面向問題[1]。本次研究所採用的資料為變遷調查宗教組中以下五項資訊：(1)訪問時間：訪問前調查資訊「訪問開始時間」、(2)性別：第 1 題「受訪者性別」、(3)出生年月日：第 3 題「請問您是什麼時候出生的」，(4)婚姻狀況： 1999 年第 6 題「請問您結婚了嗎？」及第 7 題「您目前是否與配偶同住？」，2004 年為第 6 題「請問您的結婚狀況是？」；(5)出生時辰：1999 年第 101 題，以及 2004 年第 71 題，題目皆為「請問您知道您出生的時間是幾點或什麼時辰?」。我們使用(2)性別、(3)出生年月日、以及(5)出生時辰可以得到受試者的基本命盤，再加上(1)訪問時間可以得到受試者在受訪時當下以及過去所有的大限盤和流年盤，最後由(4)婚姻狀況瞭解受試者當下是否曾經結過婚，以及當時婚姻狀態仍然是已婚、分居、離婚、抑或鰥寡。變遷調查宗教組問卷與本文章相關的問題內容如圖一所示。

圖一、變遷調查宗教組問卷與本文章相關的問題內容。

(a) 1999 年訪問時間、性別、出生年月日的相關問題

訪問開始時間 ____月_____日_____時_____分　　8□□ □□ □□ □□15

壹、基本狀況

16
1.受訪者性別：□(1)男　□(2)女　□

17　19
2.受訪者的居住地是：_____省 _____縣（市）_____鄉（鎮、市、區）　□□□

3.請問您是什麼時候出生的？　20　25
民國：_____ 年 _____ 月_____ 日　□□□□□□

[1]台灣地區社會變遷基本調查計劃

(b) 婚姻狀況的相關問題

6.請問您結婚了嗎？ 31
□(1)未婚 跳答 8 □(2)已婚 □(3)其他（請說明）_____ □

7.您目前是否與配偶同住？（選項不唸！） 32
□(1)與配偶同住 □
□(2)離婚或分居
□(3)配偶去世
□(4)因工作關係分住兩地，不常同住。
□(5)因其他緣故，夫妻分住兩地（不住在一起）。
□(6)其他（請說明）_____ 。

(c) 出生時辰的相關問題

101.請問您知道您出生的時間是幾點或什麼時辰? □□□25
□(1)知道（請說明）_____ 時 □(2)不知道 □(3)知道，但不願回答

變遷資料庫目前只有 1999 年度及 2004 年度有詢問到受訪者的出生時辰。兩次問卷總共有 3806 人受訪，不過願意且有填寫出生時辰的受試者僅有 1684 位(約 44%)，扣除掉其他無法分辨的變數後，樣本量為 1435 人。由於未填寫的人數具一定比例，我們以「有無填寫出生時辰」分成兩組，觀察這兩群受訪者在各個變量上是否有顯著的差異，如表一所示。其中性別比約 1:1，(女性略多)；未填出生時辰組的年齡顯著稍長；出生月及日則看不出兩組具顯著差異，且分布比例差異不大。

在婚姻狀況上兩組不論是「是否已婚」還是「是否已離婚」皆不具有顯著差異，就整體而言已婚者的比例約一半，離婚者約佔已婚者中的 1%左右。

表一、是否有填出生時辰之資料概況表

是否有填出生時辰	否 (N=2122)	是 (N=1684)	P-value	總和 (N=3806)
年齡				
平均值（標準差）	43.4 (15.4)	40.7 (13.8)	<0.001	42.2 (14.8)
中位數 [最小值, 最大值]	42.0 [19.0, 94.0]	40.0 [19.0, 90.0]		41.0 [19.0, 94.0]
出生年				
平均值（標準差）	47.1 (15.2)	49.7 (13.6)	<0.001	48.3 (14.6)
中位數 [最小值, 最大值]	48.0 [-1.00, 74.0]	51.0 [3.00, 74.0]		49.0 [-1.00, 74.0]

是否有填出生時辰	否 (N=2122)	是 (N=1684)	P-value	總和 (N=3806)
出生月				
1	191 (9.0%)	160 (9.5%)	0.9	351 (9.2%)
2	180 (8.5%)	142 (8.4%)		322 (8.5%)
3	154 (7.3%)	148 (8.8%)		302 (7.9%)
4	144 (6.8%)	120 (7.1%)		264 (6.9%)
5	164 (7.7%)	122 (7.2%)		286 (7.5%)
6	137 (6.5%)	126 (7.5%)		263 (6.9%)
7	152 (7.2%)	134 (8.0%)		286 (7.5%)
8	174 (8.2%)	129 (7.7%)		303 (8.0%)
9	175 (8.2%)	151 (9.0%)		326 (8.6%)
10	198 (9.3%)	148 (8.8%)		346 (9.1%)
11	179 (8.4%)	149 (8.8%)		328 (8.6%)
12	172 (8.1%)	148 (8.8%)		320 (8.4%)
遺漏值	102 (4.8%)	7 (0.4%)		109 (2.9%)
出生時				
1	0 (0%)	208 (12.4%)	NA	208 (5.5%)
2	0 (0%)	101 (6.0%)		101 (2.7%)
3	0 (0%)	154 (9.1%)		154 (4.0%)
4	0 (0%)	253 (15.0%)		253 (6.6%)
5	0 (0%)	188 (11.2%)		188 (4.9%)
6	0 (0%)	109 (6.5%)		109 (2.9%)
7	0 (0%)	186 (11.0%)		186 (4.9%)
8	0 (0%)	83 (4.9%)		83 (2.2%)
9	0 (0%)	86 (5.1%)		86 (2.3%)
10	0 (0%)	129 (7.7%)		129 (3.4%)
11	0 (0%)	111 (6.6%)		111 (2.9%)
12	0 (0%)	76 (4.5%)		76 (2.0%)

是否有填出生時辰	否 (N=2122)	是 (N=1684)	P-value	總和 (N=3806)
遺漏值	2122 (100%)	0 (0%)		2122 (55.8%)
是否已婚				
0	1040 (49.0%)	841 (49.9%)	0.528	1881 (49.4%)
1	1080 (50.9%)	837 (49.7%)		1917 (50.4%)
遺漏值	2 (0.1%)	6 (0.4%)		8 (0.2%)
是否已離婚				
0	799 (37.7%)	614 (36.5%)	0.321	1413 (37.1%)
1	18 (0.8%)	20 (1.2%)		38 (1.0%)
遺漏值	1305 (61.5%)	1050 (62.4%)		2355 (61.9%)
性別				
0	1083 (51.0%)	789 (46.9%)	0.0095	1872 (49.2%)
1	1037 (48.9%)	893 (53.0%)		1930 (50.7%)
遺漏值	2 (0.1%)	2 (0.1%)		4 (0.1%)
出生日				
1	96 (4.5%)	75 (4.5%)	0.955	171 (4.5%)
2	84 (4.0%)	61 (3.6%)		145 (3.8%)
3	46 (2.2%)	55 (3.3%)		101 (2.7%)
4	48 (2.3%)	54 (3.2%)		102 (2.7%)
5	71 (3.3%)	66 (3.9%)		137 (3.6%)
6	57 (2.7%)	55 (3.3%)		112 (2.9%)
7	59 (2.8%)	51 (3.0%)		110 (2.9%)
8	76 (3.6%)	50 (3.0%)		126 (3.3%)
9	51 (2.4%)	48 (2.9%)		99 (2.6%)

是否有填出生時辰	否 (N=2122)	是 (N=1684)	P-value	總和 (N=3806)
10	113 (5.3%)	95 (5.6%)		208 (5.5%)
11	53 (2.5%)	43 (2.6%)		96 (2.5%)
12	71 (3.3%)	50 (3.0%)		121 (3.2%)
13	62 (2.9%)	48 (2.9%)		110 (2.9%)
14	51 (2.4%)	47 (2.8%)		98 (2.6%)
15	84 (4.0%)	66 (3.9%)		150 (3.9%)
16	62 (2.9%)	50 (3.0%)		112 (2.9%)
17	56 (2.6%)	36 (2.1%)		92 (2.4%)
18	67 (3.2%)	52 (3.1%)		119 (3.1%)
19	52 (2.5%)	44 (2.6%)		96 (2.5%)
20	101 (4.8%)	101 (6.0%)		202 (5.3%)
21	55 (2.6%)	50 (3.0%)		105 (2.8%)
22	60 (2.8%)	49 (2.9%)		109 (2.9%)
23	55 (2.6%)	48 (2.9%)		103 (2.7%)
24	68 (3.2%)	42 (2.5%)		110 (2.9%)
25	82 (3.9%)	70 (4.2%)		152 (4.0%)
26	68 (3.2%)	50 (3.0%)		118 (3.1%)
27	45 (2.1%)	34 (2.0%)		79 (2.1%)
28	69 (3.3%)	60 (3.6%)		129 (3.4%)
29	53 (2.5%)	45 (2.7%)		98 (2.6%)
30	60 (2.8%)	55 (3.3%)		115 (3.0%)
31	27 (1.3%)	25 (1.5%)		52 (1.4%)
遺漏值	120 (5.7%)	9 (0.5%)		129 (3.4%)

此次分析資料 1999 年總樣本數為 1925 筆，2004 年為 1881 筆，合計 3806 筆，排除不理解題意、不知道答案、不願意回答的受訪者後，1999 年為 630 筆(約 32.73%)，2004 年為 805 筆(約 42.80%)，合計 1435 筆。最後之 1435 筆可透過排

盤軟體將受訪者的出生時辰轉換為紫微斗數命盤，進行第三章所介紹之統計分析與機器學習模型建構。

二、命理原則的量化

命理量化研究的一個重要步驟是將論命流程「標準化」以及「量化」。在這個步驟中，我們在專業命理師的指導之下，試圖把兩個部分進行量化 :(1)量化古書中的推命理論：我們將紫微斗數經典書籍中所有關於如何用命盤預測命主何時會結婚的相關敘述轉譯為可操作型定義，並換算成若干個「結婚參數」 (可以是二元變項、有序變項、或是連續變項)；以及(2)量化命理師的論命方法：根據命理師實際在論命時如何整合這些古書上的理論原則，包含命盤中各星的位置以及不同星與宮位之間的互相影響，統整成標準的定義規則與流程，並換算成一個「結婚傾向分數」。「結婚參數」以及「結婚傾向分數」都會與資訊工程師合作寫成程式，這樣一來只要輸入命主的八字、性別、以及論命時間，就可以不用經過命理師的推命得到若干個「結婚參數」以及一個「結婚傾向分數」，接著便可以和命主的真實結婚狀態進行資料分析。下章將會詳述資料分析的詳細過程。

在此先以紫微斗數其中一種算法當作本研究設計中如何將古書中的推命理論與命理師的論命方法轉譯為「結婚參數」以及「結婚傾向分數」的例子：

1. 結婚參數 1(X_1)：「命主在受訪時間以前，是否有紫微星坐於任何一次的大限盤夫妻宮」的狀態。若「命主在受訪時間以前，有紫微星坐於任何一次的大限盤夫妻宮」，則 $X_1=1$；若「命主在受訪時間以前，不曾有紫微星坐於任何一次的大限盤夫妻宮」，則 $X_1=0$。
2. 結婚參數 2(X_2)：將以上原則中紫微星換成天府星。即「命主在受訪時間以前，是否有天府星坐於任何一次的大限盤夫妻宮」的狀態。若「命主在受訪時間以前，有天府星坐於任何一次的大限盤夫妻宮」，則 $X_2=1$；若「命主在受訪時間以前，不曾有天府星坐於任何一次的大限盤夫妻宮」，則 $X_2=0$。
3. 結婚參數 3(X_3)：同 2，將天府星換成天梁星。
4. 結婚參數 4 至結婚參數 14 (X_4~X_{14})：同 2，將天府星換成沒化忌的貪狼星、有遇到天姚紅鸞的武曲星、辰戌以外天機星、左右昌曲魁鉞等六吉星，男命的太陰星、以及女命的太陽星。
5. 結婚參數 15 至結婚參數 24 (X_{15}~X_{24})：同 2，將天府星換成七殺、破軍、廉貞、天相、辰戌位的天機、四煞(火星、鈴星、擎羊、陀羅)、化忌的貪狼。
6. 結婚參數 25 至結婚參數 27 (X_{25}~X_{27})：同 2，將天府星換成地空、地劫、截空。
7. 結婚參數 28 至結婚參數 30 (X_{28}~X_{30})：同 2，將天府星換成紅鸞、天喜、天姚。

根據以上的三十項原則，我們轉換成 (X_1~X_{30})三十個結婚參數。

接著，我們請命理師將平常慣用的論命方法，使用以上三十個結婚參數描繪成一個方程式。舉例來說：命理師認為有結婚的預兆的命理跡象為 X_1~X_{14}，因此這些都加權為 1；不利結婚的預兆的命理跡象為 X15~X_{24}，加權為-1；而針對不重要的乙級星和丙級星(X_{20}~X_{30})，命理師相對不採用，則加權為 0。根據廟、旺、平、落、陷分別加權 5,4,3,2,1 的數字；同時考慮到三方四正，則將所有使用以上原則的分數都加權 50%，而把夫妻宮帶換成夾宮的兩個宮位，再重新計算兩個以上的分數，分別加權 10%；而把夫妻宮帶換成對宮的兩個宮位，再重新計算以上的分數，分別加權 30%。用如此算法得到的「結婚傾向分數」為 Z。

以上所介紹的變項(X_1~X_{30}, 以及 Z)都會與工程師合作寫成程式軟體，如此一來只需要輸入生辰八字、性別以及受訪時間，就可以自動輸出以上的準則，不需要再經由命理師針對所有受訪者一一進行推命。

三、婚姻狀態的標注與推命

我們將被預測的事件命主的「真實婚姻狀態」標示為被解釋變數 Y，為二元變量；若受訪者曾經結過婚或當下為已婚的狀態則標示為 1，未曾結過婚則標示為 0。另外在最後一個研究我們也想要探索將命理師推命的原則變成標準化流程以後，是否與他們平日的論命方式相同，因此，我們定義命理師根據命盤預測命主在填問卷之時的婚姻狀態為$\hat{Y}$。

四、與過去量化研究設計的差異

先前本團隊已經提出了兩種針對術數量化的研究設計[2]，這次的研究方法和先前的差別有以下幾點：(1)命理師的角色在先前的研究是受試者，也就是看命理師推算命運和真實發生事件的符合程度。但是在這次的研究設計中，命理師的角色是諮詢者，主要是指導研究者如何將各家流派的推命理論量化並且標準化為可操作型定義。由於命理流派彼此相差甚大，即便流派相同的命理師在論命時也常因個人經驗而可能有不同的論命結果，因此在「以命理師為受試者」的研究設計下，即便效度的結果不理想，也很容易歸因於該位命理師個人問題，而非推命數學理本身的適用性。(2)此研究設計下聚焦在命理對於「命主真實婚姻狀態」的預測，而非把所有社會變遷資料能夠推命的變項拿來預測。主要是因為過去曾以 30 人的小型資料進行分析測試，發現在婚姻狀態的預測上符合度極高，同時婚姻狀態相較於其他變項，諸如「與家人的關係」和「事業是否平順」，相對有標準的定義，評估方式較為客觀，同時婚姻對於社會的穩定性、人口、經濟、文化發展具有重要的影響，是對於解釋社會意義中較為重要的主題。

---

[2] (1)命理學量化研究之模型與方法論。林聖軒、孫保羅(民 107)。第七屆【中華傳統術數文化學術研討會】論文集。(2)中華傳統術數之量化指標發展與信效度分析。林聖軒、劉政猷、鄭志明(民 110)。第十屆【中華傳統術數文化學術研討會】論文集。

參、統計分析方法

根據上個章節的量化方法，每位受訪者皆可以得到三十個結婚參數(X_1~X_{30})，一種推命法可以得到一個「結婚傾向分數」(Z)，命理師根據命盤預測命主在填問卷之時的婚姻狀態(預測婚姻狀態，$\hat{Y}$)，以及真實婚姻狀態(Y) ，皆為結構化資料。其中除了 Z 為連續變量以外，其他皆為二元變量。我們接著就可以使用以下介紹的統計分析方法，將上述結構化資料進行量化研究。我們主要分成三個部分：(1)使用單因子分析探索三十個結婚參數與真實婚姻狀態的相關性；(2)使用單因子分析探索結婚傾向分數與真實婚姻狀態的相關性；(3)使用各種機器學習模型建構三十個結婚參數預測真實婚姻狀態的預測模型。

一、單因子分析探索三十個結婚參數與真實婚姻狀態的相關性

首先我們使用單因子分析探索三十個結婚參數與真實婚姻狀態的相關性。由於$X_1 \sim X_{30}$及Y為二元變量，可每次單獨選擇$X_1 \sim X_{30}$任何一個變量與 Y 使用卡方檢定探索其關聯性。其虛無假說為兩變量互為獨立；若 p-value < 0.05，則稱為兩變量有顯著相關。舉例來說，如果 X_1 與 Y 的卡方檢定為顯著，則代表受訪時間以前任何一次一大限盤中紫微星坐夫妻宮與婚姻狀態在此資料上有顯著的相關。這樣的結果也可以推論為：「紫微星坐大限夫妻宮」有統計上的證據能夠預測結婚狀態。

當確定變量間具有相關性後，可以再透過 Kappa 統計量來進行解釋。Kappa 統計量的值是介於-100%到 100%之間。如果是 0%的話，代表兩變量毫無關連；如果是-100%代表兩變量有高度負相關。100%則代表高度正相關。一樣取 X_1 與 Y 作為例子，若 Kappa 值為 80%，則代表受訪時間以前任何一次一大限盤有紫微星落在夫妻宮能解釋 80%的真實婚姻狀態。

在這個步驟中，會依照上述檢驗 X_1 與 Y 的相關性的方法檢驗 X_2~X_{30} 與 Y 之間的相關性，便可以知道三十種紫微命盤對於婚姻狀態的推命法則，是否在變遷調查數據中具有統計上的意義。

本文中只是簡單舉例三十種婚姻狀態的推命法則，相信有更多更精準的推命數，也都能夠用同樣的方法來進行單因子的分析，來提供各種推命法則在資料實證分析上的證據。

二、使用單因子分析探索結婚傾向分數與真實婚姻狀態的相關性

其次我們使用單因子分析探索命理師論命法則標準化後得到的「結婚傾向分數」對於真實婚姻狀態的相關性。由於結婚傾向分數(Z)為連續型變量且 Y 為二元變量，因此我們可透過將 Z 與和 Y 做單因子羅吉斯迴歸模型如下：

$$\ln(\frac{P}{1-P}) = \beta_0 + \beta_1 Z$$

其中 ln 為自然對數函數，P 代表 Y=1 的機率，β_0為截距，β_1為斜率。β_1可以被解釋為Z與 Y 相關性的指標。根據β_1的 p-value 是否顯著，可判定婚姻傾向分數Z是否能解釋真實婚姻狀態(Y)：若 Z 的 p-value 值顯著，則代表命理師對於婚姻狀態的判讀對真實婚姻狀態有顯著相關，而β_1值則代表結婚傾向分數每上升一單位，婚姻狀態的勝算比(Odds Ratio，定義為兩個狀態中 P/1-P 的比值)上升的幅度。

同時我們還可以用婚姻傾向分數與真實婚姻狀態畫出 ROC 曲線。曲線下面積(Area Under Curve, AUC)代表該婚姻傾向分數能夠預測真實婚姻狀態的能力：AUC 越接近 0.5 代表毫無預測能力，AUC>0.7 為可接受之預測能力，AUC>0.85 代表預測能力甚佳。

三、使用各種機器學習模型建構結婚參數預測真實婚姻狀態的預測模型

接下來我們想要探測三十個結婚參數能夠預測真實婚姻狀態的能力。由於結婚參數維度較高，彼此之間對於預測又常會有相互關係(例如交互作用)，因此我們使用各種機器學習模型來建構三十個結婚參數預測真實婚姻狀態的預測模型。我們首先將資料拆分成 80%訓練集與 20%測試集，進行模型訓練與預測。有時為了避免資料過度配適（overfitting，指模型太過貼合訓練資料的情況，而當放入新的預測值時預測結果不盡理想的情況），會再將訓練集進行交叉驗證法，最常用的方法是 k 折交叉驗證法。這裡的 k 指的是將訓練集拆分成子集的個數。我們這裡將訓練集拆分五份(k=5)，每次模型僅取 4 份（k-1）進行訓練，剩餘 1 份用作驗證，所以輪流驗證完後會有五次的驗證，最後綜合五次的結果得出最終模型。

接著將訓練集的資料以機器學習的方式建立分類模型，我們在此使用(1)支持向量機、(2)隨機森林、以及(3)K-NN 三種常見的模型，其原理簡介如附錄。我們利用測試集的資料評估訓練集資料建立模型的準確度來判定一個分類模型的表現，因而選擇最佳的模型作為最後的預測模型(以下稱為「機器學習命理結婚預測模型」)。

機器學習命理結婚預測模型有許多重要的意義與研究未來方向：(1)該模型的預測表現可以代表在此資料中所有「結婚參數」對於「真實結婚狀態」整體的預測能力；(2)該模型可以計算每個「婚姻參數」對於建構預測模型的重要性，此結果可以與「單因子分析探索三十個結婚參數與真實婚姻狀態的相關性」一章所述之分析相互呼應；(3)可以將上述「婚姻參數」對於建構預測模型的重要性，以及命理師心目中每個婚姻參數所代表的推命理論的重要性比較，以探討機器學習所計算出的解釋變數重要性與命理師的判斷是否存在差異。

肆、研究限制與未來展望

根據 1999 年瞿海源教授所撰寫之〈術數流行與社會變遷〉一文中提到台灣大眾對於術數行為的相信程度，透過五年一期的台灣社會變遷基本調查的調查結果，從 1985 年到 1995 年各方法（算命、看風水、安太歲等）皆有節節升高的趨勢，尤其傳統術數這塊，出乎意料地發現受到較多理性教育薰陶的高等教育份子反而有更高的比例相信術數。原作者解釋為：在瞭解現今科學發展的極限後，對於不確定對性之事物仍渴求得到一個解答的人們會有有很大的機率尋求術數的慰藉；這樣一來術數也會進一步影響生活中科學無法提供明確答案的重大決策，諸如職涯規劃、結婚生子、購屋選擇等等。由此可知，在科學昌明的當代，中華傳統術數對於台灣社會人群的影響依然極為深遠。若能針對這些術數方法的信效度提供實證量化的研究結果，便能提供大眾在參照這些術數方法進行決策時很重要的參考依據。本文章提出的量化研究設計方法便能夠給予以生辰八字為主的推命數在進行量化研究時一個很重要的指引。

本研究在執行層面上可能會遇到一些問題與限制。(1)婚姻狀態測量誤差：這可能是因為有些人在回答問題的時候，對於婚姻狀態(例如離婚一事)不願意報告實情。(2)出生時間測量誤差：在 50、60 年代晚報戶口的情形相當普遍，身分證的出生時間往往晚於真實的生日，不過生辰八字往往是真實的(為了算命以及合婚之用)。因此某種程度相信這些有回報出生時辰的出生年月日是偏向真實的回報。(3)出生地點的時差：有些專家認為必須要使用中原標準時間進行校正，但變遷調查只有詢問當下的居住地，並沒有詢問出生地的資訊來進行校正。(4)非自然產之時辰校正：有些學家也認為因醫療或非醫療理由進行剖腹產的時辰也需要校正，不過資料庫也沒有這方面的資訊。

本文提出的研究設計是針對紫微斗數對於真實婚姻狀態的量化實證研究為主，不過變遷調查中的諸多變項，諸如對於收入、學歷、以即六親相處情形也都可以用相同的方法進行量化實證研究。對於術數能夠進行推命卻在變遷調查中沒有訪問的題目，諸如收入來源、不動產、家產狀況，以及配偶的生辰八字，未來研究者也可以和變遷調查委員會的專家學者合作，將這些未來研究欲探討之問題加入變遷調查的調查問卷中。本研究方法完全適用於紫微斗數所有流派的算法，以及所有以生辰八字為主的推命數，例如四柱推命（即子平八字）；本研究設計也可以加以改良推廣到風水地理、占卜、擇日等各項中華術數領域進行量化實證研究。期待本文能拋磚引玉，有機會與各方術數大師一同合作讓各種術數成為數據科學上可重複驗證的命題，並且融合不同的學門及分析方式，提升推命數在當代的適用性，讓中華傳統術數發揚光大。

致謝

本文感謝中研院社會所傅仰止老師、齊偉先老師在研究設計方法提供寶貴的建議，國立陽明交通大學統計學研究所碩士生郭珈妤同學在資料整理上的幫忙，也感謝黃騰旭先生給予程式撰寫實務上的諮詢。

參考期刊論文

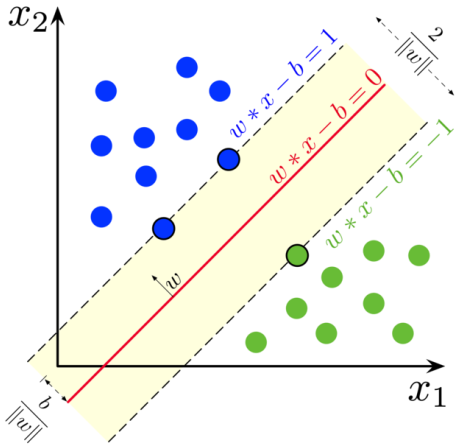
1. 瞿海源. (1999). 術數流行與社會變遷. 台灣社會學刊, 22, 1-45.
2. 命理學量化研究之模型與方法論。林聖軒、孫保羅(民 107)。第七屆【中華傳統術數文化學術研討會】論文集。
3. 中華傳統術數之量化指標發展與信效度分析。林聖軒、劉政猷、鄭志明(民110)。第十屆【中華傳統術數文化學術研討會】論文集。

附錄、本文使用的機器學習模型簡介

1. 支持向量機 (support vector machine，SVM)

支持向量機是最常用的機器學習模型之一，它的原理是使特徵空間上的間隔達到最大化的線性分類器，其分割原理是每次切分族群時使每個族群的間隔最大化，其概念如圖三所示[3]。

圖三、支持向量機(support vector machine，SVM)概念圖



以直線來說，首先紅色的線會創造兩條黑色平行於紅色線的虛線，並讓黑線平移碰到最近的一個點，紅線到黑線的距離稱為 Margin，而 SVM 就是透過去找 Margin 最大的那個紅線，來找最好的分割線[4]。

SVM 適合用來解決高維度資料的分類及迴歸問題，且當特徵數稍大於樣本數時仍可保持良好的效果（維度過大時還是需要考量資料降維的問題）。在小樣本訓練集上通常表現相較其他演算法來得好。缺點是在樣本量大時計算效率並不高、較適合用在二分類問題、對於非線性的問題很難找到適合的切割函數，且由於需要計算個樣本間的距離，對於缺失資料特別敏感。

2. 隨機森林(Random Forest)

隨機森林(Random Forest)是進階版的決策樹(decision tree)，包含多個決策樹的分類器，其輸出的類別是由個別樹輸出的類別的眾數所決定。最早是由華裔美國人何天琴於 1995 年提出[5]，第一篇論文由 Leo Breiman 於 2001 年提出[6]，通常

---

[3] File:SVM margin.png - Wikimedia Commons

[4] Yeh James (2017). [資料分析&機器學習] 第 3.4 講：支援向量機(Support Vector Machine)介紹. Medium

[5] Ho, T. K. (1995, August). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.

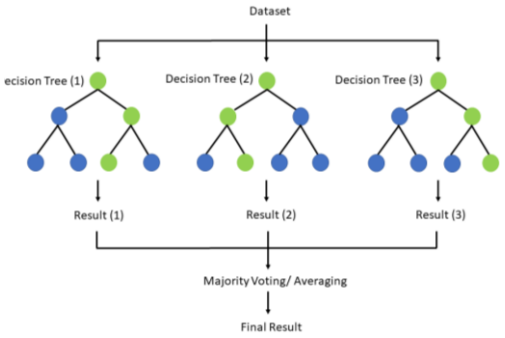
[6] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

結構上呈現如圖四所示[7]。

生成隨機森林的方法最常見的是 Bagging（Bootstrap Aggregation)。不斷從訓練集中隨機取出 N 個樣本做出決策樹，每次取出的樣本皆會再放回母體，這樣的抽樣方式被稱作 Bootstrap，是統計學上常用來進行資料估計方法。決策樹（Decision Tree）是一種利用樹形結構來進行決策的演算法，樹中每個節點代表決策，而每個分叉路徑則代表某個可能的屬性值[8]，每次新進來的資料便可根據這棵樹進行判斷，即使樹跟樹之間會有部份使用的資料重疊，訓練出的決策樹彼此間仍存在差異性，而最後透過投票方式得到最終結果。

隨機森林的優點是能處理較高維度的資料而不需先進行資料降維，並能給出哪些特徵較為重要的訊息，且相比於決策樹，較不容易發生資料過度配飾的情況。但缺點是當預測結果為連續值的迴歸問題時，它的表現不如再分類問題中表現的那麼好，且較不適合用在樣本少或者維度較低的資料。

圖四、隨機森林概念圖



3.K-近鄰演算法(k nearest neighbor，簡稱 KNN)

KNN[9]意思是取 K 個最近的鄰居，概念上就是利用「近朱者赤」的概念，也就是透過計算樣本在高維的特徵空間裡彼此的距離，假設欲預測點是 i，找出離 i 最近的 k 筆資料多數是哪一類，藉以預測 i 的類型[10]，呈現方式如圖五所示[11]。

圖五、K-近鄰演算法(k nearest neighbor，簡稱 KNN)概念圖。

---

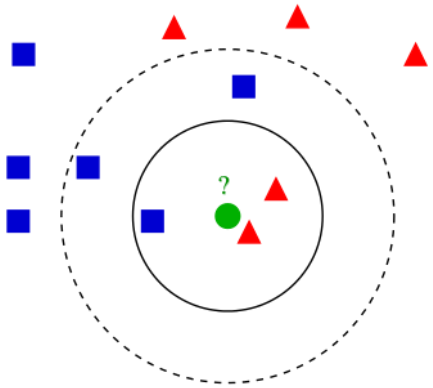
[7] What is random forest? (https://www.tibco.com/reference-center/what-is-a-random-forest)

[8] 維基百科

[9] Altman (1992). N. S. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46 (3): 175–185.

[10] 林詖北 (2017). [Machine Learning] kNN 分類演算法. Medium.

[11] File:Knn-Class.png - Wikimedia Commons



KNN 的理論簡單，既可用來做分類也可做迴歸，適用於非線性多分類問題，且理解直觀、訓練時間複雜度比大多機器學習的演算法來得低，但缺點是當資料不平衡的時候，對較稀有的類別預測準確率較低，且不如羅吉斯迴歸及隨機森林的模型具可解釋性。