論　述

# 運用行職業自動辨識系統提升統計調查效率

人力資源調查經採用「行職業自動辨識系統」以電腦自動編碼，確實提高統計調查人力效能、精進資料品質及減輕縣市人工編碼時間與負擔。本文除深入探討系統辨識與人工判定相異主因並提出改善方法，俾提升電腦自動辨識效率及推廣其應用效能。

◉ 許汶錳、侯美鈴、江文瑩（行政院主計處第4局簡任編審、科長、專員）

## 壹、前言

行（職）業分類統計向爲各界殷切需求且應用廣泛，惟因行（職）業種類眾多，若欲獲得正確的行（職）業代碼，事前需對工作人員進行行（職）業判定訓練，事後需詳細審核判別編碼資料；在行政院主計處所辦理之人力資源調查、工商及服務業普查與人口普查等作業中，行（職）業代碼之轉換，處理過程費時耗工。隨著資訊技術的進步，可先透過電腦完成前揭作業大部分工作，再針對具爭議性或模稜兩可之處進行人工判斷，以達到節省人力及時間的目的，故行政院主計處近年積極委外開發「行職業自動辨識系統」，並持續研究提升其辨識效率。

本文以94年11月至98年5月人力資源調查、95年工商及服務業普查及97年人口及住宅普查第1次試驗調查資料，運用「行職業自動辨識系統」進行行職業自動辨識結果分析，深入探討系統辨識與人工判定編碼不同之原因並提出改善方法，俾提升電腦自動辨識效率，減輕人工判定編碼之負擔及提升統計調查效能。

## 貳、系統發展沿革與簡介

### 一、系統發展沿革

為推動普查光學文字辨識（OCR），行政院主計處自民國87年與工研院電通所合作研究開發「文數字自動辨識系統」，期能有效掃描辨識大量調查資料，以節省人工輸入時間及訪問表儲存空間。初期先以人力資源調查訪問表進行開發及測試；自民國89年起陸續運用於人口及住宅普查（600萬家）、農林漁牧業普查（80萬家）與工商及服務業普查（100萬家），均節省大量資料輸入時間。為提高該系統之中文辨識功能，91年底再委外開發「多專家文字辨識組合技術」，建置行職業中文詞庫，並先針對人力資源調查訪問表上填寫之行業進行代碼轉換及程式開發，當時之行業2碼自動辨識與人工判定編碼相符比率約為70％。

為進一步提升行職業自動辨識正確率，減輕縣市調查同仁負擔，行政院主計處自92年起結合前述中文辨識資訊技術委外開發「行職業自動辨識系統」，期藉由電腦自動編碼取代傳統人工作業，有效節省行職業判定審核作業時間；該系統經多次測試，於94年11月正式完成並推行各縣市運用。

為使系統操作更符合使用者需求，於96年12月全面將系統介面由Dos版本轉換至Window平台，使用上更具親和力；另為增進系統效能，亦持續蒐集各縣市意見及配合行職業標準分類修訂，增修詞庫內容，提升行職業自動辨識效率。

## 二、系統簡介

「行職業自動辨識系統」是運用光學文字辨識（OCR）、語辭斷句（Word Segmentation）、索引分類（Indexing Identification）等技術，將文字經過語辭斷句找出關鍵索引文字，再以「中華民國行職業標準分類」為分類標準（後續加入新興產業及較口語化之詞句），將文字轉換為行業及職業代碼。如：行業——「理髮服務美容院」，系統將「理髮」及「美容院」視為語辭斷句，作為關鍵索引文字，對照「中華民國行業標準分類」搜尋行業分類內容（含子目）與語辭斷句相符之代碼，如「9620理髮及美容院」、「9690其他個人服務業之瘦身美容院」，其中符合「理髮」與「美容院」字句者為「9620理髮及美容院」，系統將視「9620」為轉換之行業代碼；職業辨識亦是如此，如「簡單記帳 會計助理」，系統將「記帳」、「會計」及「助理」視為語辭斷句，做為關鍵索引文字，對照「中華民國職業標準分類」搜尋職業分類內容（含子目）與語辭斷句相符之代碼，如「4121會計及簿記佐理員——記帳員」、「3603會計及有關助理專業人員」，其中「4121會計及簿記佐理員——記帳員」符合「記帳」及「會計」關鍵索引詞，「3603會計及有關助理專業人員」符合「會計」與「助理」關鍵索引詞，系統會以最先符

合之關鍵索引詞代碼視爲轉換之職業代碼。

本系統其他功能，如：「代碼維護」可讓使用者隨時更新行職業代碼詞庫；「詞庫維護」可將系統建置之詞庫匯出供其他調查使用者運用或產生新建詞庫或對系統詞庫進行備份工作；「設定」爲將辨識輸出檔設定至指定位址（或指定欄位），可供使用者便於進行審核工作或其他用途。其系統作業流程列於圖1。

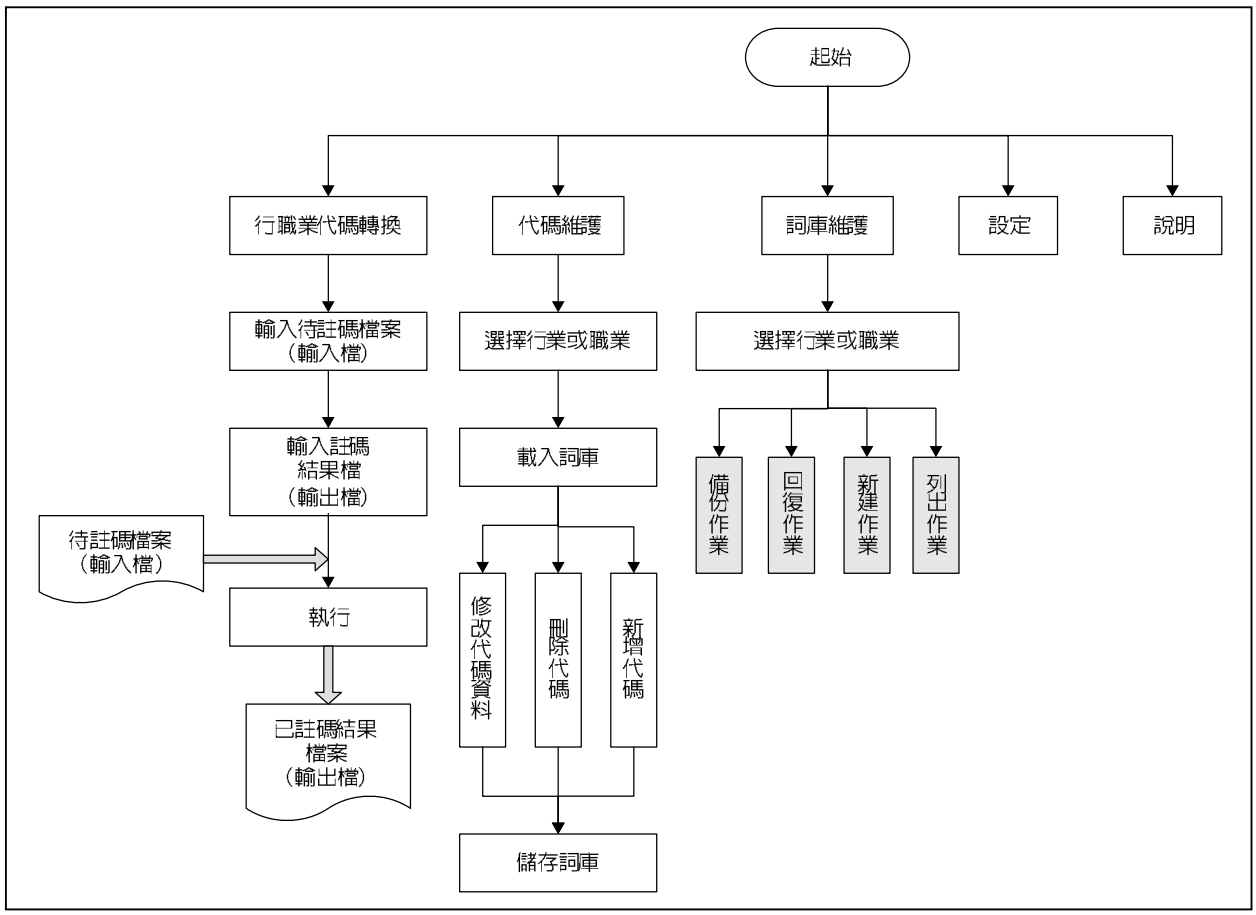
## 參、行職業電腦自動辨識結果分析

爲評估「行職業自動辨識系統」之辨識效果，本文分別就人力資源調查、95年工商及服務業普查及97年人口及住宅普查第1次試驗調查三種調查資料，運用本系統辨識後之結果進行比較分析。

### 一、人力資源調查

爲提高統計調查人力效

圖1　行職業自動辨識系統作業流程



能、精進調查品質、降低調查成本及順應各國辦理政府統計調查之技術發展，「人力資源調查」率先推動電腦輔助面訪調查作業（即由訪問員將訪查結果直接輸入電腦），自94年11月起逐步推行各縣市採用「電腦輔助面訪調查作業系統（CAPI）」，達成「統計調查工作e化，提升統計調查效率」之預定目標。為節省行職業人工判定編碼、審核之人力及時間，乃積極推廣23縣市政府使用「行職業自動辨識系統」。至96年底，所有縣市均已使用該系統進行行職業代碼轉換，對減輕縣市人力負擔與降低錯誤成效卓著。

建置完整之詞庫可大幅提升辨識效果，初期將94年11月及12月人力資源調查資料運用「行職業自動辨識系統」採93年建置之詞庫進行代碼轉換，其行職業2碼辨識與人工判定編碼相符比率僅約7成。考量93年行職業詞庫資料建置已久，不敷新興行職業使用，故分別於94年、96年及97年陸續對詞庫做全面性檢討，刪除部分贅字及增修關鍵字，並納入新興行職業，期能增加系統辨識正確率。經測試後，98年人力資源調查行業及職業2碼自動辨識與人工判定編碼相符比率約可提升至9成2與8成1（詳圖2）。

## 二、工商及服務業普查

工商及服務業普查對象係全體工業及服務業企業，普查問項包括企業之生產產品或經營服務項目，若將95年縣市普查表資料901,679筆，使用本系統97年建置之詞庫進行行業自動辨識，其2碼辨識結果與人工判定編碼相符比率可達90.83％；多數縣市甚可高達9成2至9成5（詳表1）。

經探討部分縣市行業自動辨識與人工判定編碼相符比率未達9成原因，可歸納如下：（一）都會區縣市之新興行業較多，系統詞庫未即時更新；（二）部分縣市書寫之工作內容未符系統辨識規則，導致辨

圖2　94-98年人力資源調查資料行職業自動辨識（2碼）與人工判定編碼相符比率

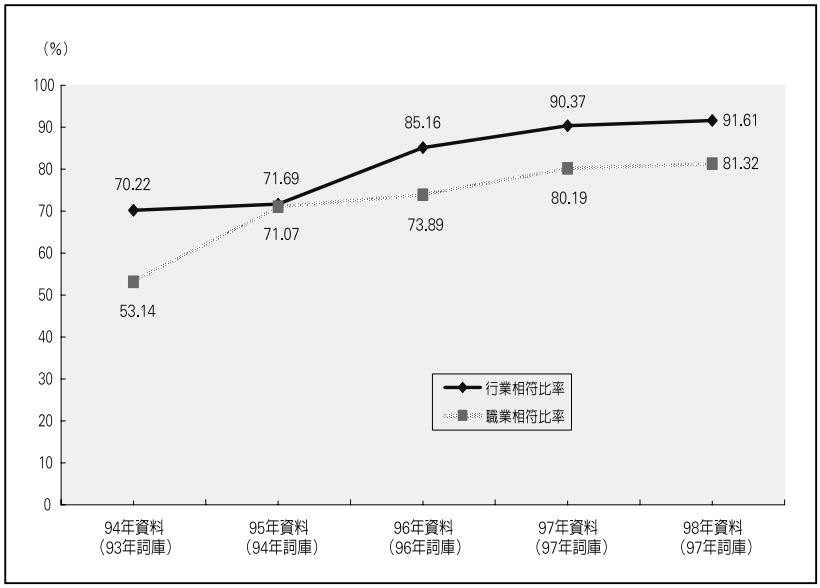


表１　各縣市95年工商及服務業普查表資料行業自動辨識（2碼）結果表

| 縣市別 | 筆　數 | 與人工判定編碼相符比率(%) | 縣市別 | 筆　數 | 與人工判定編碼相符比率(%) |
|---|---|---|---|---|---|
| 臺北市 | 135,572 | 85.79 | 高雄縣 | 50,925 | 91.74 |
| 高雄市 | 79,413 | 90.01 | 屏東縣 | 32,075 | 94.24 |
| 臺北縣 | 68,128 | 87.37 | 臺東縣 | 9,070 | 93.20 |
| 宜蘭縣 | 6,408 | 90.90 | 花蓮縣 | 13,898 | 92.10 |
| 桃園縣 | 75,213 | 92.85 | 澎湖縣 | 3,428 | 93.93 |
| 新竹縣 | 16,627 | 89.31 | 基隆市 | 14,446 | 95.07 |
| 苗栗縣 | 19,003 | 93.67 | 新竹市 | 19,194 | 91.65 |
| 臺中縣 | 74,729 | 92.88 | 臺中市 | 63,416 | 91.39 |
| 彰化縣 | 55,645 | 92.33 | 嘉義市 | 16,187 | 95.32 |
| 南投縣 | 19,466 | 92.55 | 臺南市 | 42,961 | 91.23 |
| 雲林縣 | 23,815 | 93.10 | 金門縣 | 572 | 89.16 |
| 嘉義縣 | 17,759 | 92.65 | 連江縣-馬祖地區 | 2,330 | 92.49 |
| 臺南縣 | 41,399 | 91.58 | 總　計 | 901,679 | 90.83 |

註：本統計表係採用97年建置之詞庫進行行業自動辨識。

識相符比率較低。

## 三、人口及住宅普查第1次試驗調查

若以97年11月人口及住宅普查第1次試驗調查資料6,429筆進行行職業2碼自動辨識，經與人工判定編碼比較，其中行業相符比率爲87.24％，職業爲65.81％；若改辨識至4碼，則行業相符比率降爲72.83％，職業相符比率48.72％。

表２　人口及及住宅普查第1次試驗調查資料行職業自動辨識結果表

| 縣市別 | 筆 數 | 與人工判定編碼相符比率（%） | | | |
|---|---|---|---|---|---|
| | | 行業2碼 | 職業2碼 | 行業4碼 | 職業4碼 |
| 臺北市 | 1,403 | 77.12 | 56.31 | 65.79 | 42.84 |
| 高雄市 | 1,479 | 83.77 | 63.69 | 70.99 | 45.64 |
| 臺北縣 | 627 | 86.92 | 67.78 | 73.21 | 45.93 |
| 高雄縣 | 655 | 86.87 | 66.41 | 68.55 | 44.73 |
| 基隆市 | 656 | 89.33 | 66.62 | 80.03 | 52.44 |
| 臺中市 | 894 | 89.49 | 73.83 | 79.31 | 57.49 |
| 嘉義市 | 715 | 92.17 | 75.80 | 79.30 | 58.32 |
| 總　計 | 6,429 | 87.24 | 65.81 | 72.83 | 48.72 |

註：同表1。

探討其原因，主要係該訪問表之行業填寫方式爲先填寫「場所名稱」，後填寫「主要產品或業務名稱」；職業爲先填寫「工作部門」、「職位」，後填「工作內容」。經「行職業自動辨識系統」辨識上述資料，會先對「場所名稱」或「工作部門」、「職位」產生關鍵索引字，再對「主要產品」或「工作內容」產生關鍵索引字，系統會依序將產生之關鍵索引字與詞庫進行比對，導致誤判情形偏高，故辨識效率較人力資源調查及工商及服務業普查爲低（詳表2）。

## 肆、問題分析及改進方法

### 一、問題分析

經由前述辨識結果分析可知，訪問員書寫行職業內容之順序或填寫方式爲造成「行職業自動辨識系統」辨識結果與人工判定編碼不一致之主因，茲將易導致系統誤判之處及正確作法歸納如下：

**（一）產品或經營項目、經辦工作內容未填寫清楚**：應將產品或經營項目、工作內容表達清楚，如從事電子零件製造業者，應將其主要產品項目填寫清楚，如晶圓製造或液晶顯示器製造；會計人員應將其工作內容填寫完整，如「主辦會計」或「簡單記帳 助理會計」等。

**（二）書寫方式未符辨識規則**：如從事製作鐵屋者，應填寫「鐵屋製作及架設」（屬製造業）或「搭建鐵皮屋」（屬營造業）；由於學校老師或舞蹈老師歸屬不同之職業，故應將其服務單位或工作類型詳述於前，如「小學老師」、「補習班老師」等。

**（三）填寫之產品或經營項目超過2項**：如汽車修理兼販賣或飲料批發零售業者，應擇其場所單位附加價值最大者作爲受訪者從事之工作。

**（四）書寫方式多贅字**：如從事「家教」或「務農」者，訪問員多會書寫爲「至學生家輔導功課」或「至雇主家從事農業種植工作」等（詳表3、4）。

另外，調整訪問員行職業書寫內容之順序，即將主要產品及主要職務內容書寫於前，公司名稱書寫於後（如電子公司積體電路製造，修改爲積體電路製造 電子公司），亦有助於提升辨識效率。若調整人力資源調查行職業內容書寫順序並將贅字刪除，則行業2碼辨識與人工判定編碼相符比率可由9成2提升至9成5；職業2碼相符比率亦由8成1提升至8成5（如圖3、4）。

### 二、改進方法

**（一）加強訓練訪問員書寫方式**：蒐集彙整錯誤填寫案例通報各縣市，加強訓練訪問員行職業書寫方式，要求訪問員書寫力求簡明扼要，工作場所之主要產品或業務及經辦工作內容應敘述清楚，以正確歸類行職業。

**（二）強化系統功能**：目前「行職業自動辨識系統」之資料庫維護及使用界面等功能已近完善，未來將針對系統部分功能加強改進，如結合受訪者基本特性（如：「年齡」、「教育程度」）及行職業資料進行判讀辨識，以提升行職業辨識正確率。

## 表3　行業自動辨識系統誤判歸類一覽表

| 訪問員填寫案例 | 系統自動辨識之行業 | 正確書寫方式 | 正確行業 | 備　註 |
|---|---|---|---|---|
| **（一）產品或經營項目未填寫清楚** | | | | |
| 電子零件製造 | 26-27 電子業 | 晶圓製造<br>液晶顯示器製造 | 26 電子零組件製造<br>27 電子產品製造業 | 製造業應詳細填寫工作場所主要生產之產品項目，如晶圓製造、液晶顯示器製造等，不可僅書寫行業名稱（如:電子零件製造）。 |
| 冰飲販賣 | 45-47 批發零售業 | 冰飲攤、泡沫奶茶店<br>果汁批發<br>飲料零售 | 56 餐飲業<br>45 飲料批發<br>47 飲料零售 | 販賣方式不同，行業歸類不同，故應詳列清楚工作場所主要經營或服務項目，如經營冷飲店或從事罐裝飲料批發或零售。 |
| **（二）書寫方式未符辨識規則** | | | | |
| 鐵工廠　蓋屋 | 25 製造業 | 搭建鐵皮屋<br>鐵屋製作及架設 | 41 建築工程業<br>25 金屬結構及建築組件製造業 | 僅搭建鐵屋歸營造業；製作兼搭建歸製造業，應填寫清楚。 |
| 推土機出租 | 77 租賃業 | 附駕駛推土機出租<br>拖吊車出租 | 43 專門營造業<br>77 租賃業 | 營造設備租賃附操作員者，歸營造業，未附者歸租賃業，故若出租之設備附駕駛者應註明；運輸設備租賃亦同。 |
| **（三）填寫之產品或經營項目超過2項** | | | | |
| 水果批發零售 | 45 批發業 | 水果批發<br>水果零售 | 45 水果批發<br>47 水果零售 | 生產或經營項目超過2項時，應依場所單位附加價值最大者填寫。 |
| 汽車修理販賣 | 95 維修業 | 汽車修理<br>汽車零售 | 95 個人及家庭用品維修業<br>48 汽車零售業 | |
| **（四）書寫方式多贅字** | | | | |
| 至學生家輔導功課 | 85 教育服務業 | 家教 | 96 家事服務業 | 書寫方式應力求簡明扼要，以利歸類。 |
| 至雇主家從事農業種植工作 | 43 營造業<br>95-96 個人服務業 | 務農 | 01 農、牧業 | |

## 表4　職業自動辨識系統誤判歸類一覽表

| 訪問員填寫案例 | 系統自動辨識之職業 | 正確書寫方式 | 正確職業 | 備　註 |
|---|---|---|---|---|
| **（一）經辦工作內容未填寫清楚** | | | | |
| 會計 | 41事務人員 | 主辦會計<br>簡單記帳　助理會計 | 36會計及有關助理專業人員<br>41辦公室事務人員 | 應將所從事之主要工作內容敘述清楚，以利歸類，勿僅填寫職位名稱。 |
| 綜理業務 | 12綜理業務 | 綜理業務　老板<br>推銷業務<br>販賣商品 | 12企業負責人<br>34財務及商業服務助理專業人員<br>53模特兒、售貨員及展售說明人員 | 企業負責人若有實際參與生產或服務工作，應依其實際工作內容填寫，不可僅書寫「老板」或「綜理業務」。 |
| 機操工 | 81-82設備操作工 | 玻璃設備操作工<br>印刷機操作工 | 81固定生產設備操作工<br>82機械操作工 | 應寫明操作何種機器設備，以利歸類。 |
| **（二）書寫方式未符辨識規則** | | | | |
| 老師 | 23-33 | 國小老師<br>外語補習班老師<br>運動教練、舞蹈老師<br>照顧兒童生活 安親班老師 | 23教師<br>33教學及有關助理專業人員<br>39其他助理專業人員<br>51個人服務工作人員 | 各類老師歸類不同，應將其服務單位或工作類型詳述於前，如國小老師、舞蹈老師，以利歸類。 |
| 櫃檯服務生 | 42接待人員 | 收銀員　電動玩具店<br>餐飲服務生<br>售貨員　便利商店 | 42顧客諮詢接待事務人員<br>51個人服務工作人員<br>53模特兒、售貨員及展售說明人員 | 應將所從事之主要工作內容詳述完整，如餐飲服務生或商店售貨員或櫃檯收銀員等。 |
| 體力工 | 91-92體力工 | 鋼管搬運工　機械公司<br>建築泥沙搬運工 | 91送件工、搬運工及有關體力工<br>92生產體力工 | 應寫明何種性質工作之體力工（如送件搬運工或營造、建築體力工等），以利歸類。 |

**（三）賡續增建詞庫**：近年來新興行職業變化大，為能即時反應現況，將賡續擷取95年工商及服務業普查及按月人力資源調查資料進行詞庫增建

工作，並建立隨時更新機制；另爲增廣本系統之運用層面，將擴增行職業代碼由2碼辨識增爲4碼且提升其辨識正確率，期推廣至人口及住宅普查及工商及服務業普查，以提升本系統應用效能。

圖3　98年2月人力資源調查變更行業書寫方式之辨識（2碼）結果與人工判定編碼相符比率

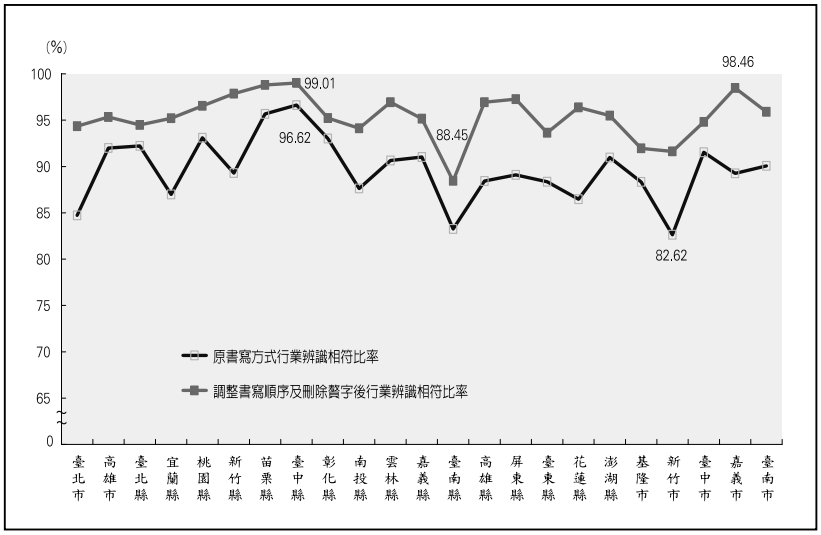
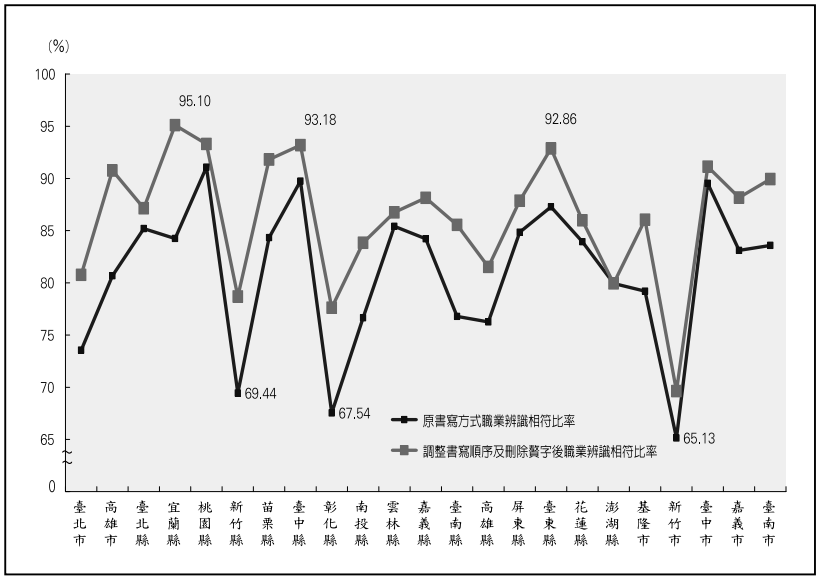



圖4　98年2月人力資源調查變更職業書寫方式之辨識（2碼）結果與人工判定編碼相符比率




## 伍、結語

人力資源調查運用「行職業自動辨識系統」以電腦進行行職業自動編碼以來，有效節省人工判定編碼作業時間，大幅減輕調查人員工作負荷及提升資料品質。爲提升「行職業自動辨識系統」之效率，經研究發現，改變行職業中文填寫順序及隨時增修詞庫，可提高行職業2碼辨識正確率至9成5及8成5。未來除擴增行職業註碼由2碼至4碼外，將再加強訪問員行職業書寫內容訓練，建立隨時更新詞庫及強化系統功能機制，期提升行職業辨識正確率，並推廣至各項普查及抽樣調查應用。❖