郵件投遞成功率提升一最佳路徑規劃

|台大資管B隊|:孫君傳、劉正宇、簡辰安

|指導老師|:魏志平教授、孔令傑教授

摘要

- **目的**: 提高郵件**投遞成功率, 減少重複投遞成本**以增加郵務士及郵局整體工作效率
- 投遞成功率提升方案:
 - 找出影響投遞成功率的因素: 資料視覺化分析與統計檢定
 - **獲得最佳投遞時間:**提出透過<u>預測模型</u>、<u>知識管理系統(KMS)</u>及<u>事先預約</u>的方式,讓 郵務士在投遞前了解手上郵件投遞狀況。
 - **發展最佳路線規劃:**將投遞過程描繪成VRPWT問題,配合最佳投遞時間,計算出最 佳的投遞路線。
- 成果: 透過程式驗證方案可行性, 確實提高投遞成功率以降低成本

理念一回歸原點

中華郵政有著穩固的基礎客戶與規模,也一直致 力於諸如i郵箱、民間通路、便利商店合作等創新。

但除此之外, 我們認為中華郵政可以回到原點,

透過郵務數據的分析, 提高郵件投遞的效率來:

- 縮短原先郵務士送件及善後的時間
- 解放人力資源以發展新業務

優勢

- •紙本郵件領先市場
- •員工數量多且優質
- •穩定的金融業務→ 支援大膽郵務決策

劣勢

- •易固化的組織結構
- •國營事業的公共責任

機會

- •資訊科技成熟
- •追求效率的郵務市場

威脅

•其他物流公司對於環境的快速應變

一、背景資料分析

-郵件投遞成功率與影響投遞成功率之參數

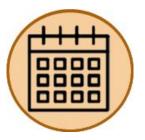
整年約18.8%的郵件會投遞失敗

18.8412% 81.1588% 其中以下因素會影響投遞成功率



雨量:

<u>雨量大到一定程度投遞</u> 成功率有顯著下降



時段, 星期幾:

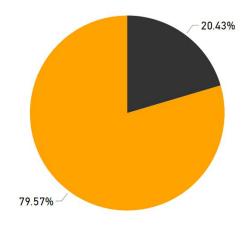
9點~11點、16~18點高機率成功 假日的成功率低於平日

分析流程

● 我們選用高雄800584局號的資料作為這次報告的依據選擇此局號的原因為其處理的 件數較多,且在這一季的運送路線較為規律。

●失敗 ●成功

- 總體1億以上郵件中,超過20%的郵件投遞失敗 (如右圖)原因將在後面分為幾個部分討論:
 - 1.投遞時段的影響
 - 2.投遞日為星期幾的影響
 - 3.投遞日天氣的影響



時段

在各個時段中的投遞成功率

右圖可發現時間是影響成功率的顯著原因

高機率成功:

9點~11點、16~18點

高機率失敗:

8點前、12~15點、19點

可能原因:

- 1.特定時段收件者比較不容易在家或無法收件
- 2.刷條碼習慣的問題,可能郵務士習慣在特定時間段刷投遞失敗的狀態



投遞日性質(星期幾)

在一個星期中各日的投遞成功率

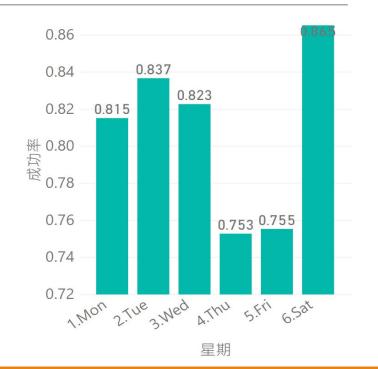
由右圖可以發現到

假日的成功率高於平日

可能原因:

一般民眾假日通常會在家中, 容易簽收郵件

星期四與五的收件成功率稍低於其他日子



天氣

天氣是影響投遞的重要因素之一,從資料中可以看出:

雨量大到一定程度, 投遞成功率有顯著下降

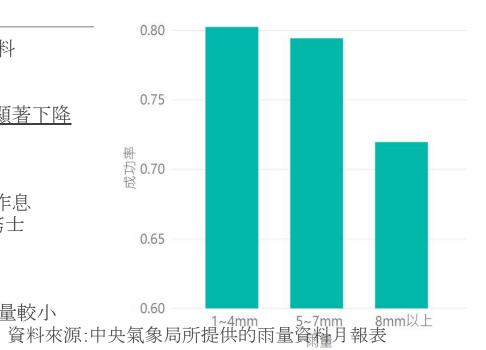
可能原因:

輕微下雨不太影響收件者與郵務士作息 當雨大到一定程度,便會影響到郵務士 遞送的難易度與客戶的行為。

備註:

此資料集取自冬季的高雄,因此雨量較小 無法討論雨量很大的情況 資料來

在各種雨量中的投遞成功率



可能影響因素

我們希望找出統計證據,以證明不同的因素確實會導致不同的寄件成功率,如此便能蒐集相關變數,並建置相對應的預測模型。

Why

懷疑時間、日期等是影響寄件成功率的因素

How

使用 Pearson's Chi-Squared test of contingency table

What

找出統計證據, 證明不同的變數會影響寄件成功率

Chi-Squared test of contingency table

不同日

Ho: 不同日的郵件寄送成功率 一樣。

H₁:不同日的郵件寄送成功率不一樣。

MON TUE WED THU FRI SAT Success 208824 234623 201855 226744 214015 16146 Fail 47348 45798 43499 74463 69320 2515

Pearson's Chi-squared test

data: (week)

X-squared = 11403, df = 5, p-value < 2.2e-16

morning afternoon evening night Success 21297 236036 844864 10 Fail 5493 242072 32238 3140

Pearson's Chi-squared test

data: (time)

X-squared = 431940, df = 3, p-value < 2.2e-16

不同時段

H₀:不同時間的郵件寄送成功率 一樣。

H₁:不同時間的郵件寄送成功率不一樣。

Chi-Squared test of contingency table

不同雨量

H₀:不同天氣(雨量)的郵件寄送成功率 一樣。 H₁:不同天氣(雨量)的郵件寄送成功率不一樣。

by 中央氣象局所提供的雨量資料月報表

1~4mm 5~7mm 8~10mm Success 953147 56971 42157 Fail 234627 14750 16423

Pearson's Chi-squared test

data: (rain)

X-squared = 2385.9, df = 2, p-value < 2.2e-16

小結:

透過Pearson's Chi-squared test,我們有充分的統計證據拒絕虛無假設,也就是**不同時間點、日期和 雨量所對應的寄件成功率並不相同**,不論兩者間是相關還是因果關係,這些變數都是我們建置預測模型的重要參考。

二. 獲取最佳投遞時間之方法

01. 預測模型

蒐集影響投遞成 供的資料,使用統 計和機械學習模 型,預測客戶收件 的最佳時間。

02. 知識系統

整合資訊資源,包括郵務士自身經驗和投遞的歷史資料,紀錄客戶收件的最佳時間。

03. 事先預約

客戶主動預約收件時間、地點,提供郵務士更彈性的收件模式和精確時間。

考慮成功收件的時間,能減少重複投遞及善後的次數,進而降低運送成本接下來詳述以上三種收集最佳投遞時間的方案

方法1: 預測模型

將投遞成功與否設為應變數,將更多可能影響的參數:日期、時間、氣溫、雨量、紫外線指數和同時間同區號的包裹數設為自變數,由於成功與否為決定性問題(是/否),我們選擇使用 logistic regression 建置模型。

Why

了解參數對成功率影響的程度,提供規劃投遞路線的依據

How

使用R langauge 建置 Logistic Regression 模型

What

預測郵件在不同時間點、星期、天氣等狀態下的投遞成功率

Logistic Regression

*女人	success	time	week	address	temp	rain *	UVL	amount
2109356	14	evening	WED	80700	19.6	1	9	234107
2109357	14	afternoon	FRI	80700	15.0	1	7	234107
2109358	H4	afternoon	THU	80200	23.7	7	8	170119
2109359	H4	afternoon	WED	80200	24.0	1	10	170119
2109360	14	evening	THU	80200	16.5	3	6	170119
2109361	14	evening	WED	80700	21.1	1	11	234107
2109362	H4	afternoon	WED	80700	15.5	4	7	234107
2109363	14	afternoon	FRI	80700	15.0	1	7	234107
2109364	H4	afternoon	THU	80700	23.5	1	7	234107
2109365	14	afternoon	FRI	80700	15.0	1	7	234107

資料來源: 800584局號中郵件狀態代碼被登記為14(成功)與H4(不成功)、且掛號號碼中有郵遞區號的投遞資料,共約97萬筆,每筆資料加上郵遞區號、當日平均氣溫、雨量、紫外線指數和同梯寄件數

時間:早上(9~11點)、下午(12~15)、傍晚(16~19)和晚上(剩餘時間)

訓練樣本: 7成的資料作訓練樣本, 剩下3成作比對

model

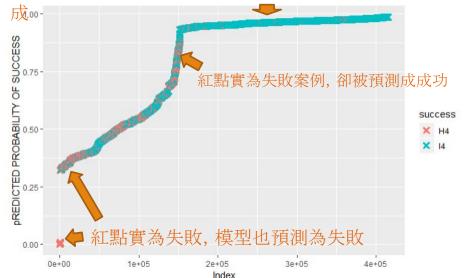
```
Deviance Residuals:
                 Median
-2.9161
         0.1958
                 0.2623
                                   3.5482
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.300e-01 1.789e-02 -18.450
timeevening 3.366e+00 6.497e-03 518.096
timemorning 1.520e+00 1.564e-02 97.174
timenight -5.604e+00
                       3.149e-01 -17.793
weekMON
            6.614e-02 8.498e-03
week5AT
            8,900e-01 2,656e-02 33.503
weekTHU
            3.676e-02 7.763e-03 4.736 2.18e-06
weekTUE
            3.836e-01 8.262e-03 46.427
            6.146e-01 8.415e-03 73.040
weekWED
           -1.748e-02 8.109e-04 -21.556 < 2e-16 ***
           -1.778e-02 1.047e-03 -16.970
            2.409e-02 1.108e-03 21.731 < 2e-16 ***
            3.025e-06 2.892e-08 104.597 < 2e-16 ***
amount
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '1
(Dispersion parameter for binomial family taken to be 1)
   Null deviance: 1402502 on 1385149 degrees of freedom
Residual deviance: 942039 on 1385137 degrees of freedom
AIC: 942065
Number of Fisher Scoring iterations: 7
```

檢查模型

```
Deviance residuals is centered at 0
Deviance Residuals:
               Median
   Min
            10
                           30
                                    мах
                                                                                                      Confusion Matrix and Statistics
-2.9161
       0.1958
                0.2623
                         0.3314
                                3.5482
                                          (good)
Coefficients:
                                                                                                                Reference
                                                                                                       Prediction
            Estimate Std. Error z value Pr(>|z|)
                                                                                                                            14
          -3.300e-01 1.789e-02 -18.450 < 2e-16 ***
(Intercept)
                                                                                                                  43693 30357
                                                   Wald's test for all coefficients:
timeevening 3.366e+00 6.497e-03 518.096 < 2e-16 ***
                                                                                                              14 41189 300305
           1.520e+00 1.564e-02 97.174 < 2e-16 ***
timemorning
                                                      all statistically significant
timenight
          -5.604e+00 3.149e-01 -17.793
                                                                                                                     Accuracy: 0.8278
weekMON
           6.614e-02 8.498e-03
                               7.783 7.07e-15 ***
                                                                                                                       95% CI: (0.8267, 0.829)
week5AT
           8.900e-01 2.656e-02 33.503 < 2e-16
                                                                                                          No information Rate: 0.7957
weekTHU
           3.676e-02 7.763e-03
                               4.736 2.18e-06 ***
                                                                                                          P-Value [Acc > NIR] : < 2.2e-16
           3.836e-01 8.262e-03 46.427 < 2e-16 ***
weekTUE
                                                                          模型有82.8%的正確
weekWED
           6.146e-01 8.415e-03 73.040 < 2e-16 ***
                                                                                                                        Kappa: 0.444
          -1.748e-02 8.109e-04 -21.556 < 2e-16 ***
temp
          -1.778e-02 1.047e-03 -16.970 < 2e-16 ***
rain
                                                                                                       Mcnemar's Test P-Value : < 2.2e-16
UVL
           2.409e-02 1.108e-03 21.731 < 2e-16 ***
           3. 5e-06 2.892e-08 104.597 < 2e-16 ***
amount
                                                           战功將成功案例預測成成功的準確度有9
                                                                                                                  Sensitivity: 0.5147
Signif. codes: 0'***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '1
                                                                                                                  Specificity: 0.9082
                                                                                                               Pos Pred Value: 0.5900
(Dispersion parameter for binomial family taken to be 1)
                                                                                                               Neg Pred Value: 0.8794
                                                                                                                   Prevalence: 0.2043
   Null deviance: 1402502 on 1385149 degrees of freedom
                                                                                                               Detection Rate: 0.1051
Residual deviance: 942039 on 1385137 degrees of freedom
                                                                                                         Detection Prevalence: 0.1782
AIC: 942065
                                                                                                            Balanced Accuracy: 0.7115
Number of Fisher Scoring iteration: 7
                                                                                                             'Positive' Class: H4
```

預測模型檢討

藍點實際為成功案例,模型預測為成功的機率值高達9



此圖為測試資料被歸類為成功的機率值,同前頁模型的 specificity所示,此模型對成功寄件的辨識度較高。 我們認為預測寄件成功率,相當於預測客戶的行為模式(寄件失敗多為送件時客戶不在),由於一季的天氣變化不大、郵遞區號所內涵的地理範圍資訊太少,甚至郵務士的行動過於規律也會影響結果,這次影響較大的變數只有寄件時間點和日期(星期)。

要能提高預測準確度,除了嘗試不同模型如:SVM, random forest和MLP等, 更重要的是讓訓練資料更具代表性。

我們建議若中華郵政希望**預測**不同時間點的寄件成功率時,應在**不違反客戶隱私權**的前提下,採用**長時間、具有客戶端的行為模式的代表性資料**,如更精確的**地理位置、去識別化用戶id、結合長期使用郵政服務**情形的統計和**推測人口結構**等資料,如此可以提高預測準確率。

預測模型結論

- 影響成功率的參數:時間>星期>天氣>郵遞區號件數(高雄冬天氣候過於穩定)
- 郵務士可以在投遞前了解個別郵件的投遞成功率。此預測模型判定為成功的郵件有高達 91%可以投遞成功。對郵務士而言,可以更專注在那些較能成功投遞的郵件,避免延後 投遞
- 可以透過模型了解如何改變投遞模式以增加投遞機率,舉例而言,可以在容易投遞成功的時間盡可能投遞郵件,並在容易失敗的時間進行其他業務,達到更好的時間分配
- 雖然現階段沒有,若**增加用戶id甚至地點資訊**,不只能加強模型的精確度,更可以進一步分析用戶或地點的特性,在正確的時間與狀態下將郵件投遞至特定位置,達到效率最大化
- 接下來會**結合最佳化模型**, 進行更具規模的應用

Logistic Regression

參數

^	success	time ÷	week	temp =	rain [‡]	UVL 0
6	14	night	TUE	18.9	0.0	5
7	14	night	TUE	18.9	0.0	5
8	14	night	TUE	18.9	0.0	5
9	H4	night	TUE	18.9	0.0	5
10	14	night	TUE	18.9	0.0	5
11	H4	night	WED	18.9	0.0	4
12	14	night	WED	18.9	0.0	4
13	14	night	WED	18.9	0.0	4
14	14	night	WED	18.9	0.0	4
15	14	night	WED	18.9	0.0	4
16	14	night	WED	18.9	0.0	4
17	H4	night	WED	18.9	0.0	4
18	14	night	WED	18.9	0.0	4
19	14	night	WED	18.9	0.0	4
20	14	night	WED	18.9	0.0	4
-			4			

資料來源:台中400583郵局中郵件狀態代碼被登記為4(成功)與H4(不成功)、且屬於**包裹或快捷**, 共約3700筆, 每筆資料加上當日平均氣溫、雨量、紫外線指數

時間:早上(9~11點)、下午(12~15)、傍晚(16~19)和晚上(剩餘時間)

訓練樣本: 7成的資料作訓練樣本, 剩下3成作比對

model

```
Deviance Residuals:
-1.8800
         0.6162
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.382222
timeevening -1.300085
                       0.114778 -11.327
timemorning 0.034616
timeniaht
           -1.870795
                       0.446145 -4.193 2.75e-05 ***
weekMON
            -0.110224
                       0.077946 -1.414 0.15733
week5AT
            -0.509367
                       0.289015 -1.762 0.07800
weekTHU
            0.095075
                       0.073892
                                 1.287 0.19821
weekTUE
            0.184052
                       0.068977
                                 2.668 0.00762 **
weekWED
            -0.009196
                       0.072893 -0.126 0.89961
            -0.014201
                       0.005551 -2.558 0.01051 *
rain
            0.003071
                       0.002882
                                1.066 0.28657
UVL
            0.034441
                       0.015536
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 12808 on 12391 degrees of freedom
Residual deviance: 12618 on 12380 degrees of freedom
AIC: 12642
Number of Fisher Scoring iterations: 4
```

預測模型結論

```
Confusion Matrix and Statistics
          Reference
Prediction
        T4 766 2900
              Accuracy: 0.7863
                95% CI: (0.7728, 0.7994)
    No Information Rate: 0.7879
    P-Value [Acc > NIR] : 0.6043
                 Kappa: 0.0279
 Mcnemar's Test P-Value : <2e-16
           Sensitivity: 0.02792
           Specificity: 0.99044
         Pos Pred Value: 0.44000
        Neg Pred Value: 0.79105
             Prevalence: 0.21206
         Detection Rate: 0.00592
   Detection Prevalence: 0.01346
      Balanced Accuracy: 0.50918
       'Positive' Class: H4
```

- 準確度:0.9048
- 影響成功率參數:**時間 > 星期 > 天氣**
- 具代表性參數不足使目前模型不易預測 會投源失敗的郵件
- 建議提升模型品質的方法:
- o 使用不同模型如: SVM, random forest和MLP等
- o 放入更精確的**地理位置、去id化用戶資訊**等參數
- 結合長期使用郵政服務 情形的統計和推測人口結構

方法2:知識管理系統(KMS)

現在許多郵務士早已自己有一套運送習慣,知道哪些客戶什麼時候較容易送到,卻缺乏一個有系統的整合,因此可以建置一個資料庫,儲存並管理這些寶貴的經驗,好處如下:

- 歷史資料可輔助完善方法1之預測模型
- 只需資料儲存及詢問郵務士,成本低廉
- 提升人力調度彈性:

爾平新手與老手的經驗差,新手上路或投遞區域更換效率不會變差, 進而可依據當日郵件總量或郵局營運情況進行人力調派



方法3:APP事先預約

- 被動收集:由收件者主動預約希望的收件時間(比起預測時間區間,更為精準)
- 衍生應用:除了預約時間,更能預約收件地點(ex:改到上班地點取貨,更為彈性)



三.最佳路徑規劃

確保郵差一次就成功投遞,進而減少重複投遞成本和增加郵差投遞效率

目標

方法

結合客戶郵址、道路資訊、過去運送成本統計和給定送貨時間,將投遞過程變成整數線性規劃 (ILP)的問題

結果

將運送過程視為

Vehicle Routing Problem with Time Windows 問題, 算出在不同時間點送達至不同客戶的最佳路線。 模型

s. t. :₽

$$\sum_{k}^{V}\sum_{j}^{N}x_{ijk}=1$$
, $\forall i\in C$

$$\sum_{j=1}^{N} x_{0jk} = 1, \quad \forall k \in V \quad .$$

$$\sum_{i}^{N} x_{ihk} - \sum_{j}^{N} x_{hjk} = 0$$
, $\forall h \in C, k \in V$

$$\sum_{i=1}^{N} x_{i,n+1,k} = 1, \ \forall k \in V$$

$$\sum_{i=1}^{C} \sum_{j=1}^{N} d_{i}x_{ijk} \leq q, \ \ \forall k \in V$$

$$\sum_{k}^{V} \sum_{j}^{N} x_{0jk} \leq |V|, \quad \forall k \in V, j \in N$$

$$s_{ik}+t_{ij}-s_{jk}\leq M_{ij}\big(1-x_{ijk}\big),\quad\forall i,j\in N,k\in V$$

$$a_i \leq s_{ik} \leq b_i, \quad \forall i \in N, k \in V$$

$$x_{ijk} \in \{0,1\}, \quad \forall i,j \in N, k \in V$$

$$s_{ik} \in N^+, \forall i \in N, k \in V_+$$

VRPWT 模型

決策變 數

$$x_{ij}^k \coloneqq \begin{cases} 1, & k_{th}$$
 郵差是否從 i_{th} 客戶端移動至 j_{th} 客戶端 o/w $s_{ik} \coloneqq k_{th}$ 郵差開始服務 i_{th} 客戶的時間 ω

決策目

Obj: ₽

$$Min \sum_{k=1}^{V} \sum_{i=1}^{N} \sum_{j=1}^{N} c_{ij} x_{ijk}$$
 總運輸成本

模型參 數

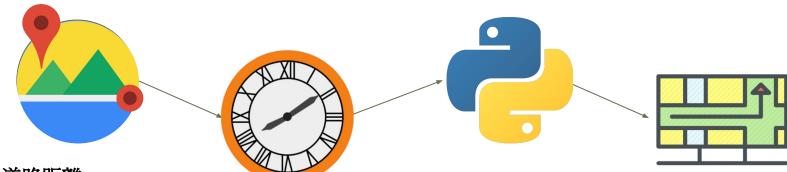
> $c_{ij} \coloneqq \text{從i}_{th}$ 客戶到 j_{th} 客戶的交通成本。 $t_{ij} \coloneqq \text{從i}_{th}$ 客戶到 j_{th} 客戶的所需時間。 $q \coloneqq \text{車輛載重/空間}$ 。 $d_i \coloneqq \text{i}_{th}$ 客戶的計件重量/體積。

 $[a_i,b_i]\coloneqq \mathrm{i}_{\mathsf{th}}$ 客戶的最佳投遞時間 |

 $M_{ij}\coloneqq \max\{b_i+t_{ij}-a_j\}$ (移動時間上限) 。

使用預測模型或 過去歷史資料 推測高機率投遞 成功的時間區間

做法



道路距離

使用google提供的 distance-matrix api

運送成本

過去運送的統計

投遞時間

透過預測模型預測、知識庫的紀錄或事前預約而得到

整數規劃

使用solver計算:





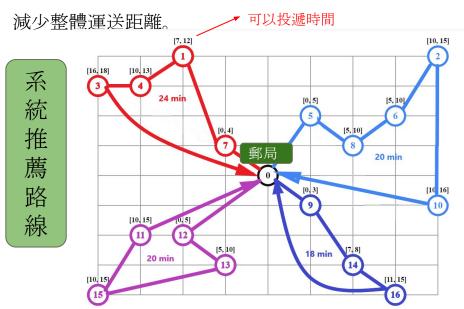
OR-tools等工具

結果呈現

得到抵達各點的時間和所屬郵差,可將結果輸出成圖表

實際範例

使用程式運行VRPTW規劃路線,可考慮到顧客有空的時間,**避免投遞失敗**,且可考慮人腦不易想出的路線,得以

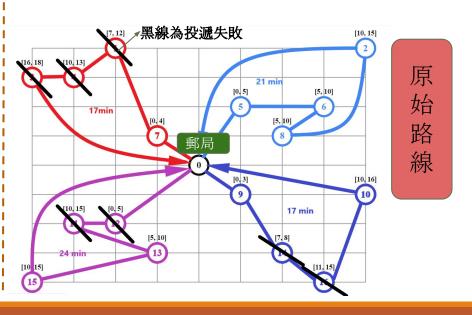


使用OR-tools計算各郵差抵達不同客戶的時間的結果

Route for vehicle 0: 0 Time(0,0) -> 9 Time(3,4) -> 14 Time(7,8) -> 16 Time(14,14) -> 0 Time(21,21) Time of the route: 21min

使用greedy演算法模擬郵差實際投遞路線

- 1. 一昧前往最近的地址, 整體路線不一定最短
- 2. 不清楚投遞時間, 會發生投遞失敗的個案



最佳路徑結論

- NP Problems

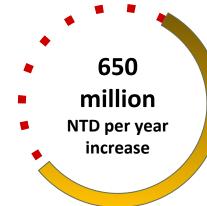
 P Problems

 NP Complete
- ▶ 參數: 地址、路況、運送成本統計、郵務士數量、最佳投遞時間
- 成果: 在指定時間內投遞的最佳路徑, 1.減少重複投遞和2.降低運輸成本
- 限制:
 - 此問題為NP-complete, 也就是說處理的數量越多, 計算過程會越複雜, 以至於使用低階配備會無法計算。
 - 對於**少量客戶**的郵遞區號(幾千戶), 可以透過solver (ex,gurobi, or-tools)**直接運算**。
 - o 如果**數量過多**, 則必須退而求其次使用heuristic algorithmn, 針對VRPTW問題, 我們嘗試使用過 Genetic algorithmn (1)、Ant colony algorithmn (2)等加入隨機變數的演算法, 已逼近最佳答案, 運算 速度會快上許多, 且答案都不錯。
 - Vehicle Routing with Time Windows using Genetic Algorithms Sam R. Thangiah Artificial Intelligence and Robotics Laboratory Computer Science Department Slippery Rock University Slippery Rock PA 16057 U.S.A.

效益評估



• 最高提升 投遞成功率 18.8% (若取得最佳投遞時間使 投遞幾乎不投遞失敗)



- 投遞失敗成本計算 失敗件數(3889萬件)
 - x 每件花費時間(2.5分)
 - x 每分鐘成本(6.73元)
 - = 654,324,250元



- 人力調度彈性效率提升、不受經驗限制
- APP可帶來**額外效益**

ex: 推廣郵局服務...

可行性分析

- 預測模型:可發展,現階段無地點、客戶id等資料,但理論可行
- 知識管理系統:可行,僅有儲藏成本和與郵務士溝通
- APP: <u>可行</u>, 建置、推廣及後續管理成本, 為配合客戶時間作業流程可能須調整
- 最佳路徑運算時間限制: <u>可行</u>, 根據說明會導師表示, 最快會在郵件投遞後的1.5小時寄出, 表示至少1.5小時以上的緩衝時間, 可根據時限調整運算資源。
- 最佳路徑系統建置成本: <u>可行</u>, 基本運算單位(電腦)和OR研究人員。
- 作業流程調製:可行,郵務士需配合新的路徑規劃來整理郵件與改變習慣路線

總結論

• 目標:

增進郵務士的投遞效率 及減少後續處理投遞失敗 的時間

• 最佳收件時間:

- 預測模型:抽取台北、台中、高雄各一郵局的資料,找出時間、星期、天等有影響力的參數,建立精準度約90%的**模型預測收件時間**,但未有地點和客戶相關參數,解釋力並不好
- 知識管理系統:可**紀錄並分析過往郵務士投遞經驗**,進而整合客戶的最佳投遞時間資訊
- 事先預約APP:由**客戶主動通知可投遞時間**,並有預約收件地點等額外應用

• 最佳路徑規劃:

給定收件時間並計算最佳路經,提高初次投件成功的機率的同時也壓低運送和時間成本,預 估最高可節省6.5億元