統計模擬期末報告

台灣交通事故分布與時間分析

110354016 蔡耀德

關鍵字:Kernel density estimation、Bandwidth selection

壹、 摘要

此報告先進行一維與二維資料分析交通數據，找出台灣交通事故熱點與時間，透過直方圖、散佈圖、熱點圖點資料，再進行同時考量地點與時間的三維資料分析，找出最佳核密度估計函數(KDE)，此報告將著重於核函數（Kernel function）與環寬(Bandwidth)的選取，並試著比較這些函數與參數的關係，最後得出合適的密度函數。

貳、 動機及背景

鑒於台灣交通事故發生率居高不下，身為機車族的我也對此感到擔心，為了解決容易發生交通事故的問題，想透過分析 A1 及 A2 交通事故之交通熱點與時間，試著找出台灣交通事故熱點與時間。

參、 研究方法與目的

以政府資料開放平臺公布之 110 年度 A1 及 A2 類(110 年 1 月-6 月、110 年 7 月-12 月)交通事故資料為依據，考量經度、緯度、發生時間等變數，主要探討以核密度估計方法找出最佳三維台灣交通事故密度函數，並分析台灣各地之事故熱點，透過選取適當的環寬以及核函數估計台灣 A1 及 A2 交通事故之三維密度函數，並評估交通情形。

肆、 本文

一. 資料處理與探索

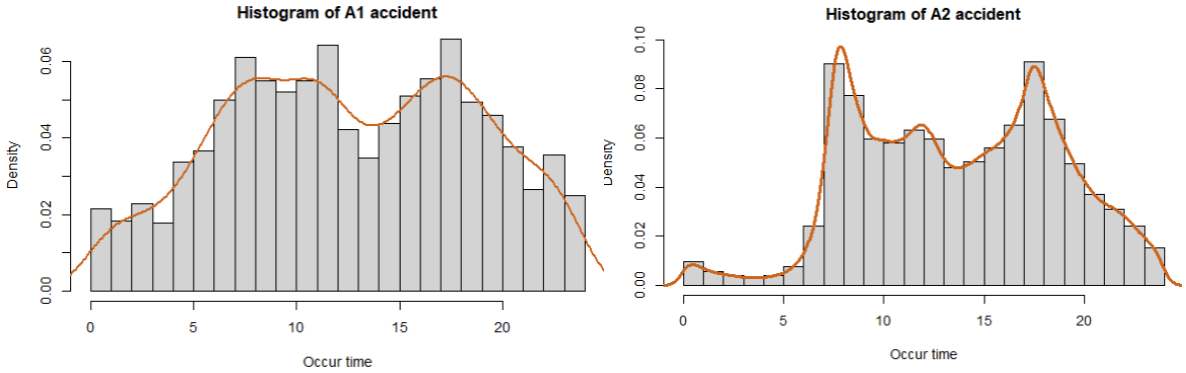
1. 預處理

先將資料中時間年份進行處理，並提取資料中事故發生日期與時間，轉為小時與月份之數值，以便分析經緯度及時段、月份的關係，並將交通事故之死傷人數分為死亡及受傷人數

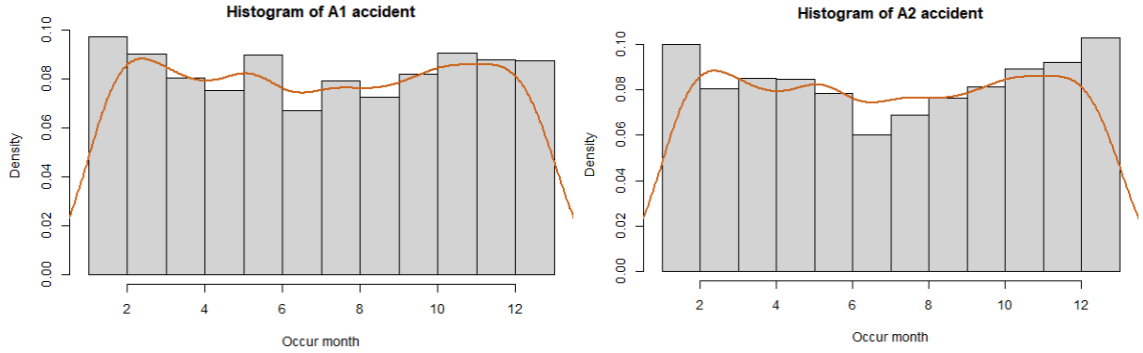
	發生時間 <chr>	經度 <dbl>	緯度 <dbl>	死亡 <int>	受傷 <int>	發生時段.時. <dbl>	發生日期.月. <dbl>
1	2021/1/1 04:49	120.5601	22.56711	1	0	4.816667	1.032258
2	2021/1/1 04:57	120.6800	24.07070	1	0	4.950000	1.032258
3	2021/1/1 07:30	121.3034	24.19183	1	0	7.500000	1.032258
4	2021/1/1 09:00	120.5367	24.25235	1	0	9.000000	1.032258
5	2021/1/1 13:32	121.4656	25.04191	1	1	13.533333	1.032258
6	2021/1/1 15:40	120.4166	23.63405	1	0	15.666667	1.032258

2. 直方圖

透過發生時間之直方圖並加上核密度估計的函數線(R 預設之常態核函數)，可以大致發現在早上 7~12 點及下午 5~6 點發生 A1 交通事故機會最高，而 A2 交通事故則是在早上 7~9 點及下午 5~6 點發生 A2 交通事故機會最高，而在凌晨間 0~5 點間 A1 與 A2 交通事故發生機率差異較為明顯。

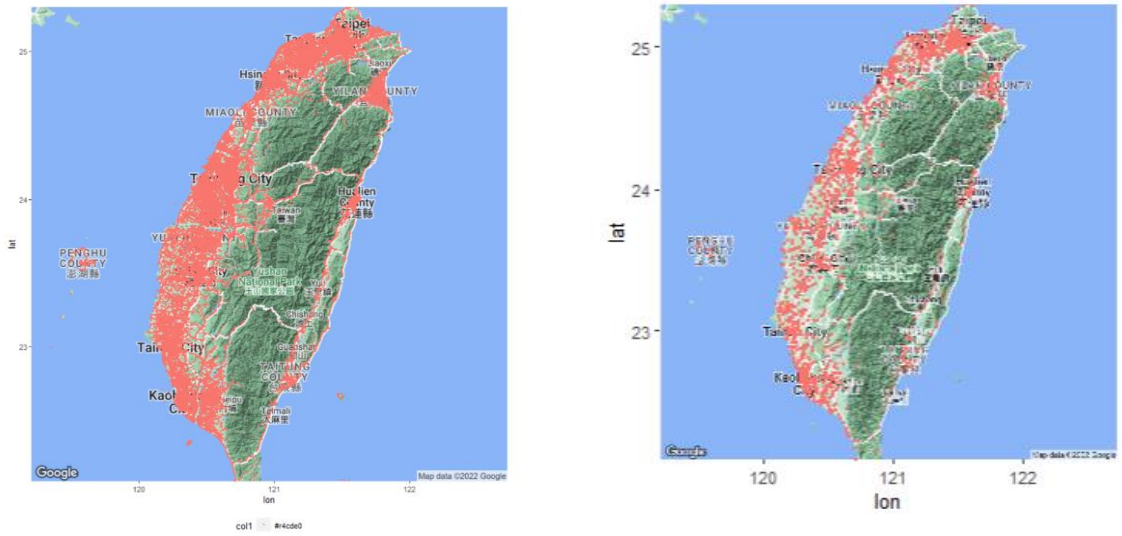


透過月份之直方圖可以看出在秋、冬季(10~1 月)發生 A1 及 A2 交通事故機會較高，而在核密度估計圖中 1 月及 12 月邊界估計值明顯較低，與核密度函數特性有關。



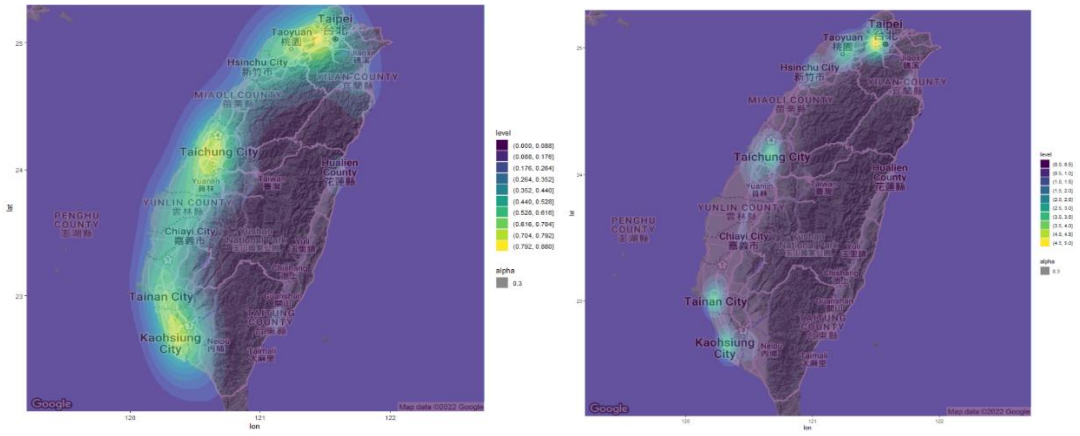
3. 散佈圖

使用 Google Maps Api 串接 R 再加上 ggplot 可以繪製一張台灣的交通事故散佈圖。下方分別為台灣 A1 及 A2 交通事故散點圖，A1 及 A2 交通事故散佈在西部縣市，分布於人群居住地。



4. 經度與緯度二維密度估計圖

依照 A1 與 A2 交通事故案發之經度與緯度，使用 R ggplot2 中 geom_density_2d_filled 函數，使用預設條件之 Gauss function 作為核函數、Grid points 選擇 20×20、並依照 bandwidth.nrd 函數選取最佳環寬可畫出兩種情形之二維核密度估計圖，由下圖可看出 A1 交通事故的資料較少因此估計值較高且廣，在北部、中部、南部地區之人口密集地發生機率較高；而 A2 交通事故資料相較之下集中於特定縣市，在五個直轄市以及桃園市與新竹市交通事故發生機率高，其中雙北地區明顯最高。

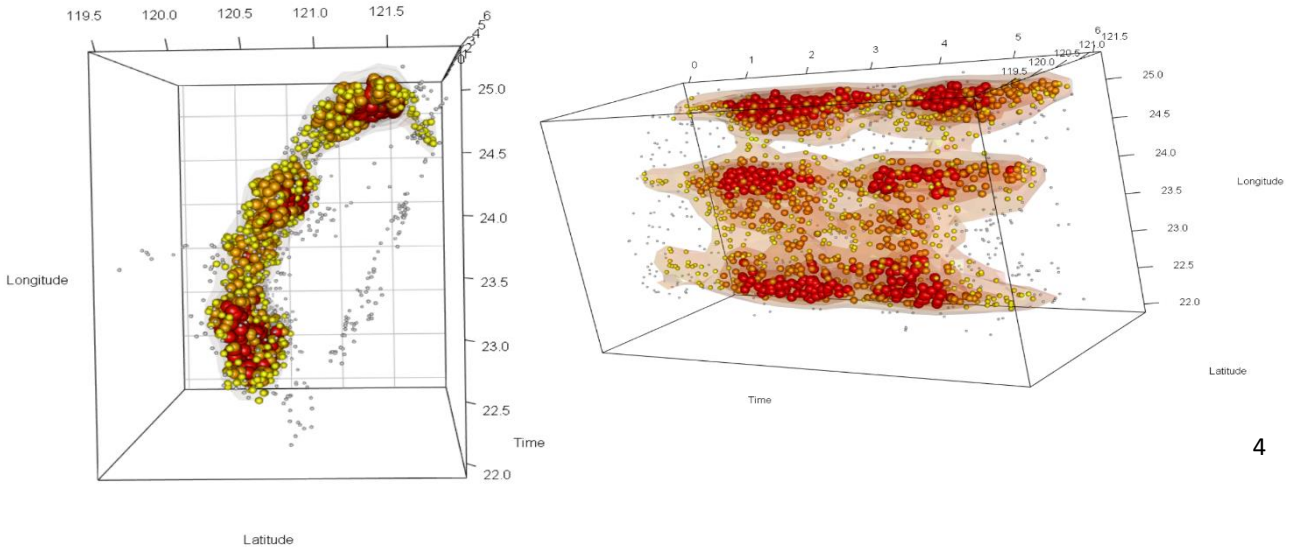


二. 核函數密度估計

在此選取常用的核函數：Uniform, Epanechnikov, Triangular , Gaussian function 及依照 Cross-Validation Methods、Plug-in Method 以及特定函數之選取方法選取適當環寬，來估計經緯度與時段(時)、經緯度與月份、經緯度與受傷人數之密度估計值，再依 3D 箱型圖的結果選取適合的函數圖形，箱型圖中 X、Y、Z 軸為經度、緯度與時間，此外為了將圖形方便呈現與解釋，將**時間除上 4** 再進行估計。

※ 3D 圖形在報告上較難呈現，在此進行圖形解釋。

X 軸為經度、Y 軸為緯度、Z 軸為$\frac{\text{時間}}{4}$，下圖正面為經度與緯度，由於交通事故發生在西部之密度明顯較高，因此分析時皆將圖形**轉為側面(時間軸)**進行分析。



1. Cross-Validation Methods

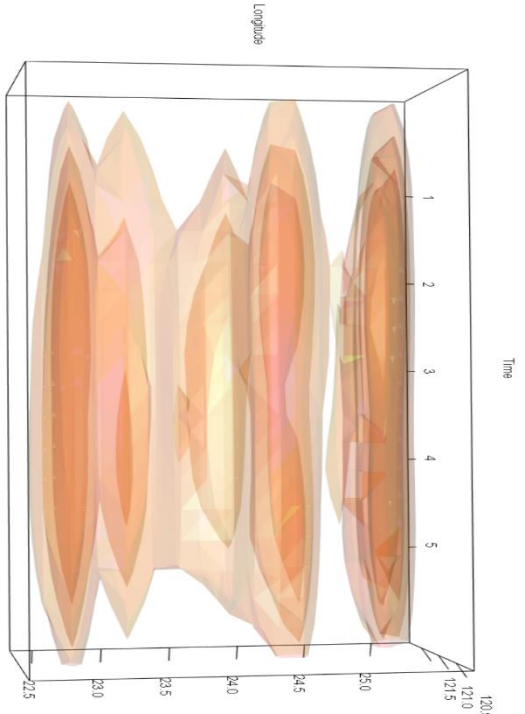
在此選取兩種交叉驗證方法 Biased Cross-Validation 以及 Unbiased Cross-Validation。

i. Biased Cross-Validation

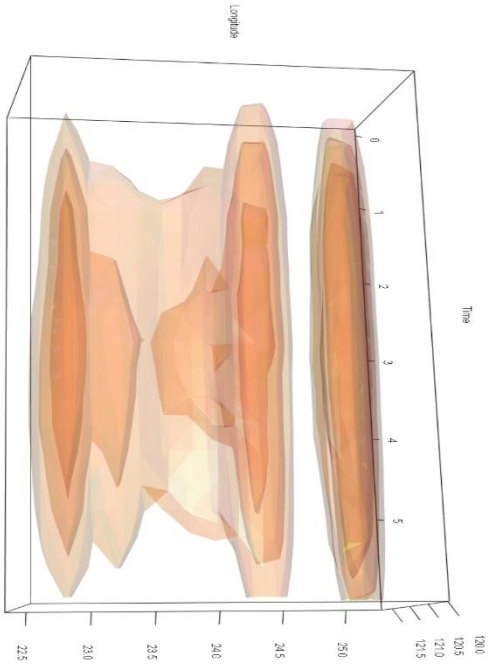
依照 bcv 函數規則，選取 Bins 的數量 1000 個，並取預設之 Lower bound 及 Upper bound 進行 10-fold Cross-Validation，得到以下圖形結果。(圖片為橫向)

可以大致看出 Epanechnikov、Triangular function 在 BCV 選取規則中估計較為清楚，而 Gaussian function 較為平滑。

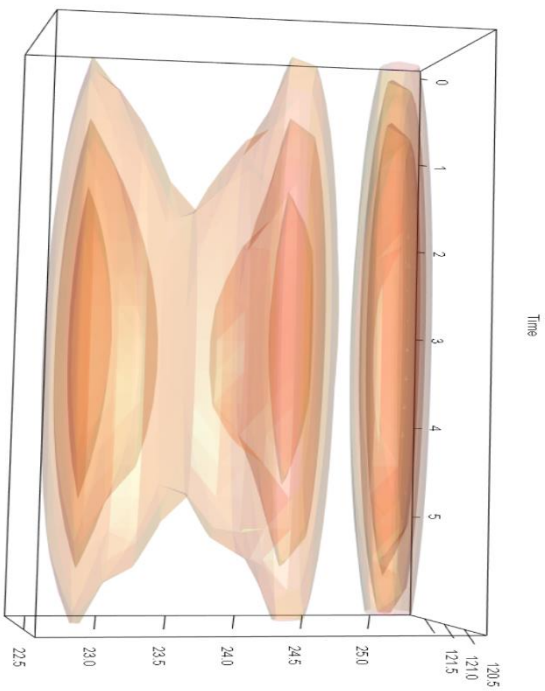
Triangular function



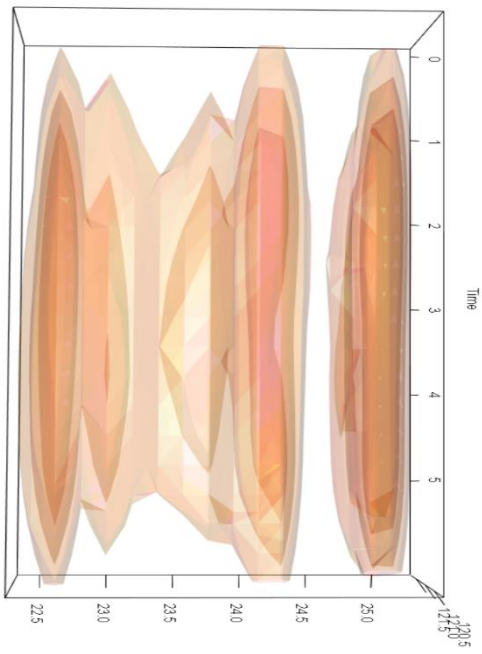
Uniform function



Gaussian Function



Epanechnikov function

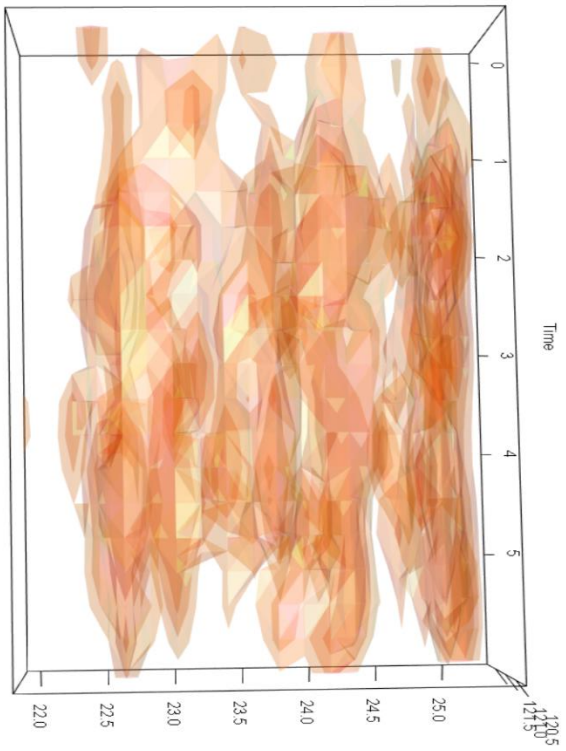


ii. Unbiased Cross-Validation

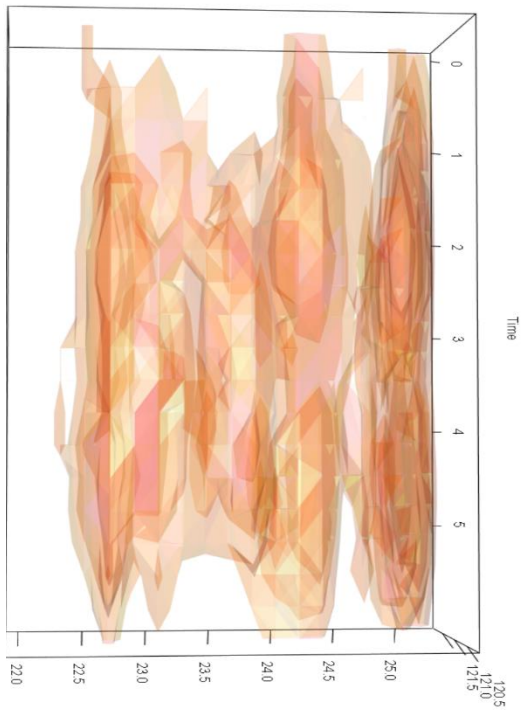
依照 ucv 函數規則，選取 Bins 的數量 1000 個，並取預設之 Lower bound 及 Upper bound 進行 10-fold Cross-Validation，可以得到以下結果。(圖片為橫向)

相較 Biased Cross-Validation，此方法密度分布層數較多，Gaussian function1 表現亦較其他三者平滑。

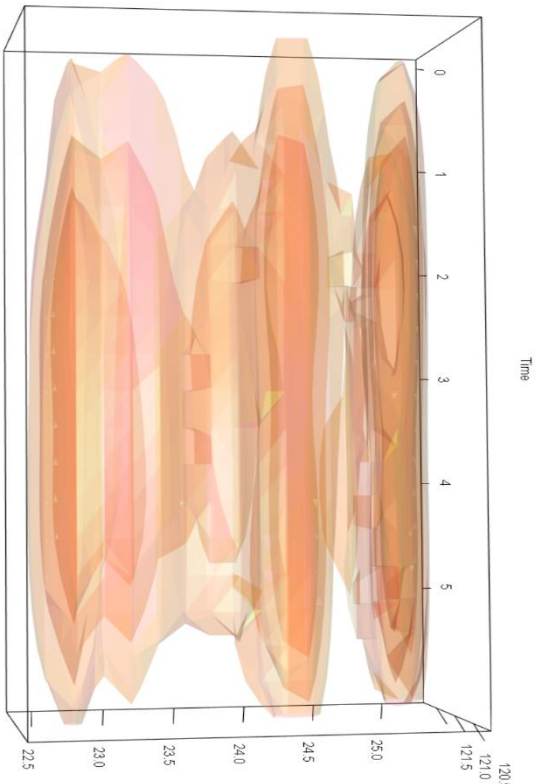
Triangular function



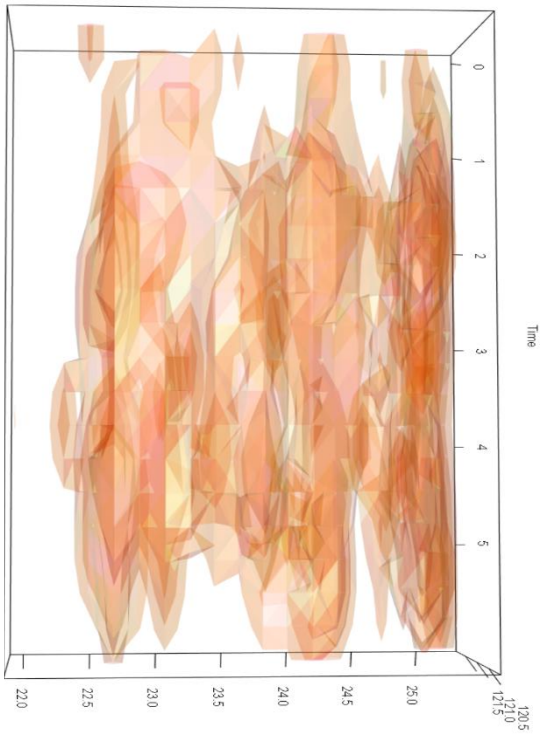
Uniform function



Gaussian Function



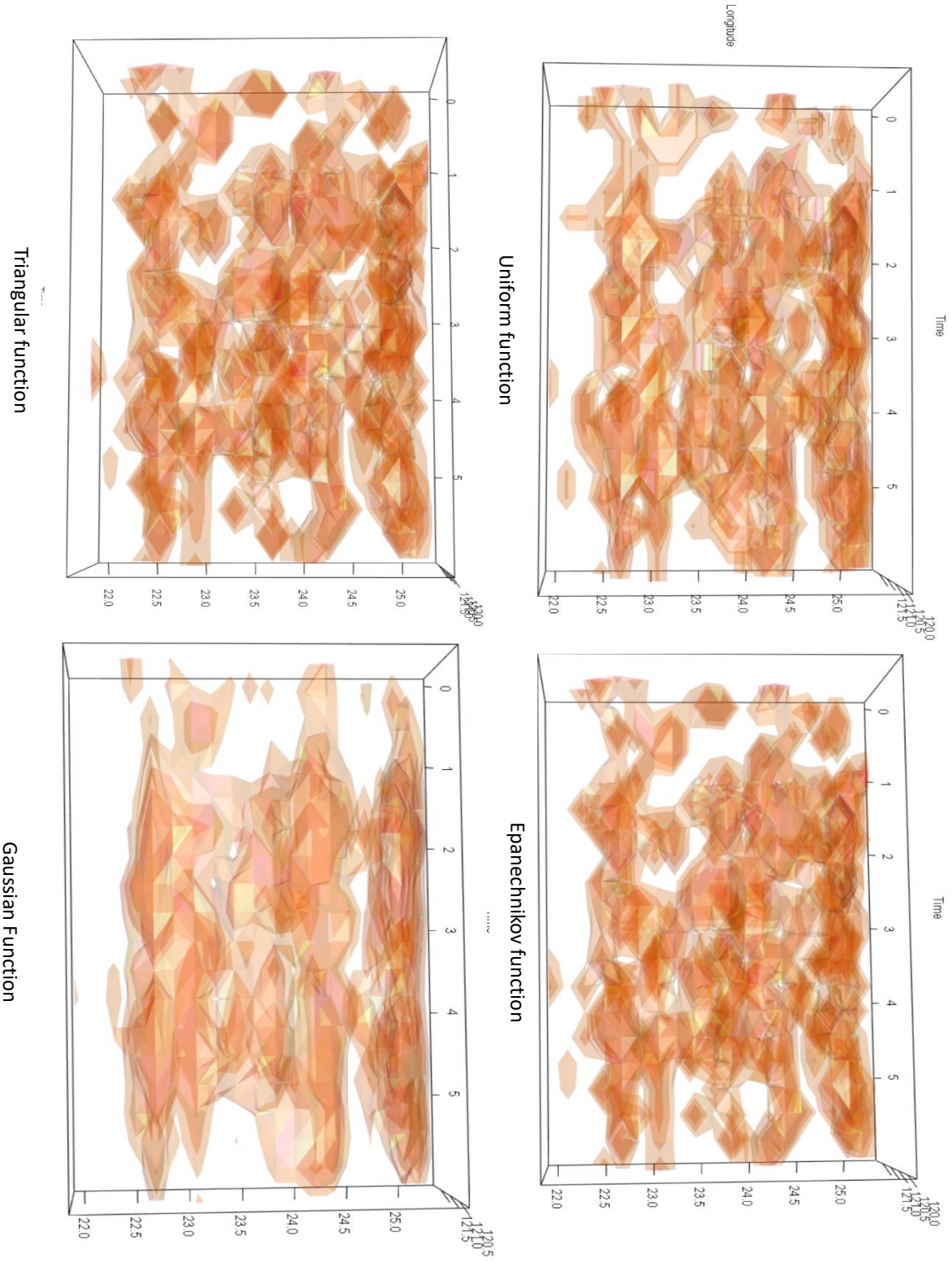
Epanechnikov function



2. Plugin Method

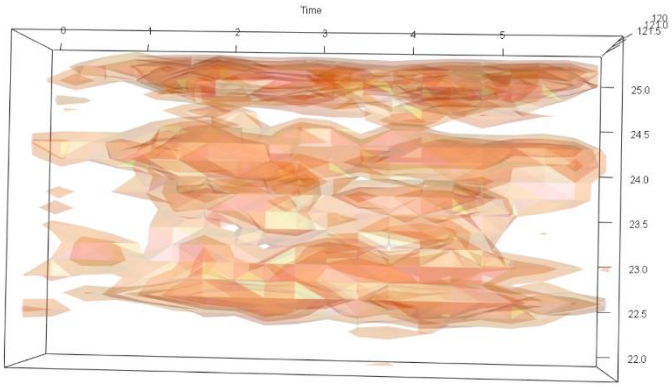
透過迭代的方法找出環寬估計 AMISE，使用 plugin.density 函數並迭代 200 次後選取環寬，Plugin method 的迭代方式相當多元，在此使用 Eva Herrmann 之方法來找出環寬。(圖片為橫向)

此方法之 Gaussian function 亦較其他三者平滑，而估計函數分布較廣。估計能力較 Cross-Validation Methods 差，看不出明顯趨勢。

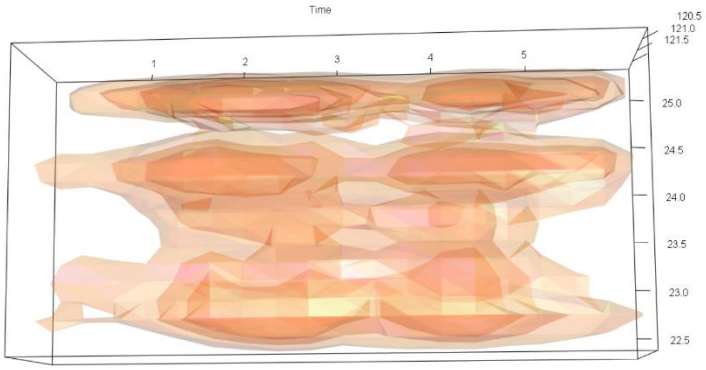


3. Other methods

在選取環寬的問題中，Silverman 在 Density Estimation for Statistics and Data Analysis 書中提到 **the Rule of Thumb**，即核函數為 Gaussian function 之最佳選取環寬為 $\hat{h} = 1.06 \times \min\left(\sqrt{Var(x)}, \frac{IQR}{1.34}\right) \times \mathrm{n}^{\frac{-1}{5}}$，其中$x$分別為 X 軸、Y 軸、Z 軸之數值，這邊使用 $\hat{h}$ 以及 R 語言中 kde3d 函數之預設環寬$\frac{2}{3}\hat{h}$，從下方圖形可知選取環寬 $\frac{2}{3}\hat{h}$ 時較 $\hat{h}$平滑。

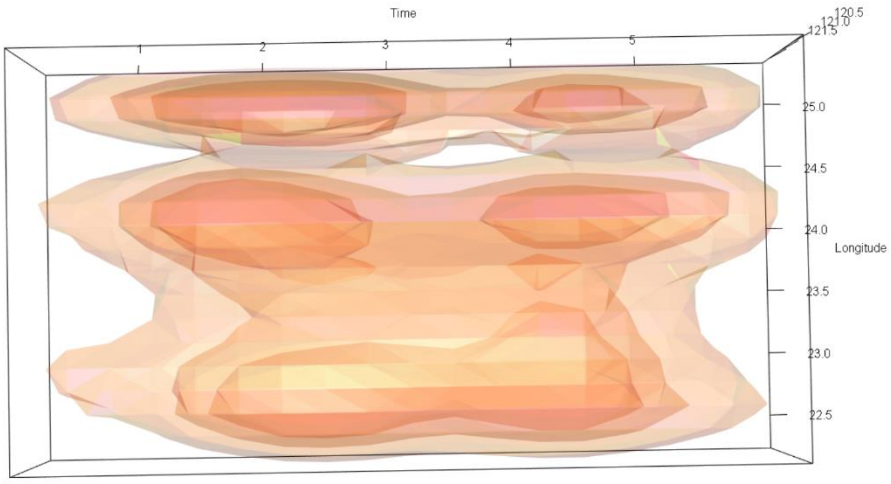


Gaussian Function with bandwidth $\hat{h}$



Gaussian Function with bandwidth $\frac{2}{3}\hat{h}$

此外，Silverman 也在同一本書中提到核函數 Triangular function 之最佳選取環寬為$\hat{h} = 2.575\sqrt{Var(x)} \times \mathrm{n}^{\frac{-1}{5}}$，其中$x$分別為 X 軸、Y 軸、Z 軸之數值，由下圖可知此方法估計密度較為平滑。



Triangular Function with bandwidth $\hat{h}$

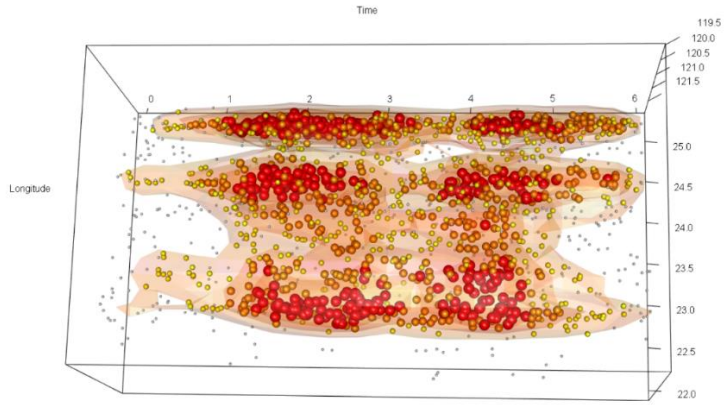
5. 結論

由上述方法中可以大致看出在相同環寬情況下，核函數間的估計結果差異不大，相較之下，相同核函數且不同環寬選取之間估計差異較為明顯，因此環寬選取的方法相當重要。David W. Scott1 與 George R. Terrell 在 1987 年論文 Biased and Unbiased Cross-Validation in Density Estimation 提到，在樣本不足的情況下 BCV 會過於平滑且 UCV 之 variance 較大，因此在 A1 交通事故資料估計能力較差，而 Plugin Method 估計能力不佳，亦不考慮。因此，在經度、緯度與時間方面，最終選擇**核函數為 Gaussian function、環寬為**$\frac{2}{3}\hat{h}$之核密度估計結果。

三. 套入模型

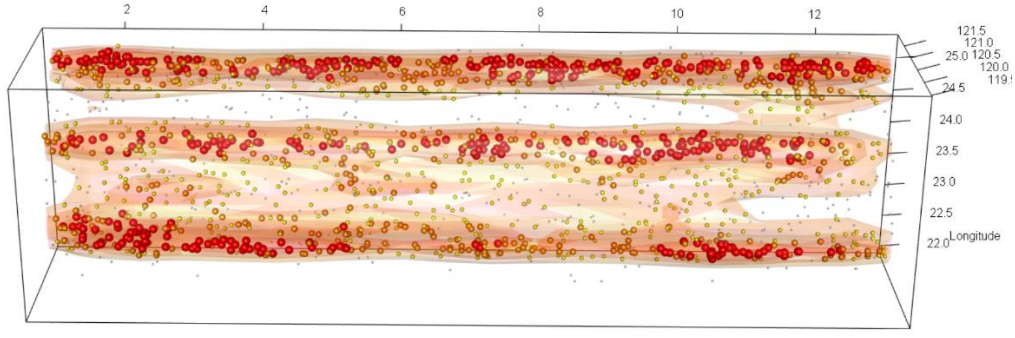
1. A1 交通事故

將各資料點之估計值代入可得以下結果，密度依照 3 個 25、50、75 百分位數將資料分為灰色、黃色、橘色以及紅色，因此可以看出北中南部在早上與下班時段發生交通事故機率高(Time 為$\frac{\text{時間}}{4}$)。



經度、緯度、時間之密度估計圖

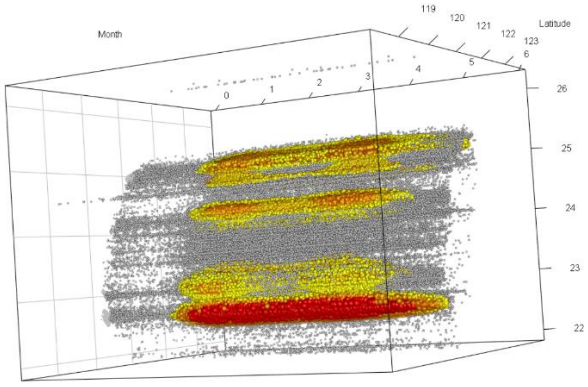
從**經緯度與月份**之估計圖可看出雙北一年四季皆為交通事故好發時段，台中區與高雄區則是集中在 10 月至隔年 5 月之間。



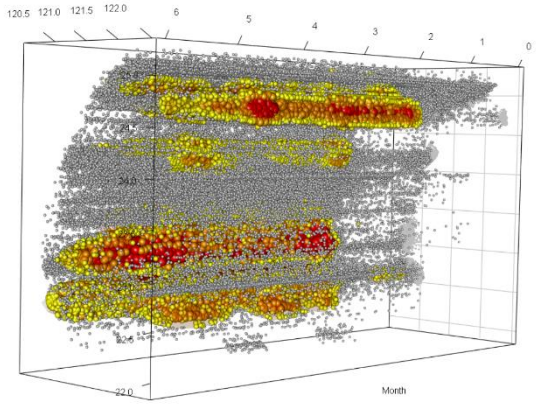
經度、緯度、月份之密度估計圖

2. A2 交通事故

依照相同規則選取 Gaussian function 且 Bandwidth 依 the Rule of Thumb 選取，可看出交通事故集中於雙北、台中、台南及高雄，其中雙北、台中、台南交通事故集中在早上以及下班時段，而高雄則從早上至晚上交通事故密度最高。

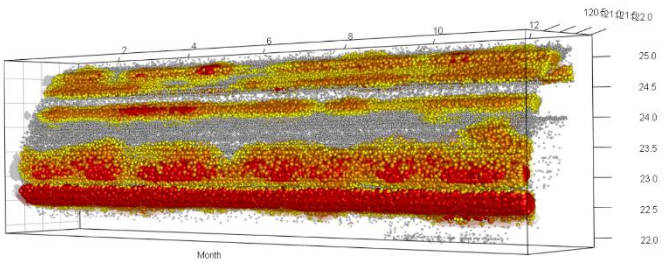


經度、緯度、時間之密度估計圖(正面)

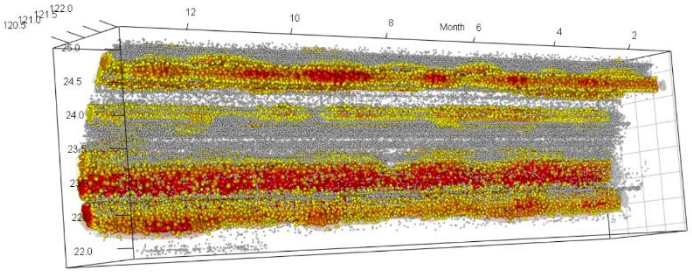


經度、緯度、時間之密度估計圖(背面)

從下圖正面與背面可知，台南與高雄在每月交通事故密度最高，其次為雙北、桃園、台中。



經度、緯度、月份之密度估計圖(正面)



經度、緯度、月份之密度估計圖(背面)

伍、 結語

A1 交通事故中，好發時段為早上至中午、下午下班時段，且雙北一年四季皆為交通事故密度皆相當高，台中區與高雄區則是集中在 10 月至隔年 5 月之間。

A2 交通事故時間上主要集中在高雄且在時段上無明顯差別，其次為台南、台中及台北，而高雄、台南交通事故月分皆相當密集，與原先一維與二維分析結果相近。

在核密度估計結語中提到 BCV 在樣本不大時估計結果會過於平滑，因此在 A1 交通事故資料(1600 筆)估計效果並不好，然後因 A2 交通事故資料足夠大量(35 萬筆)，因此估計結果也相當不錯，因此亦可使用 BCV 選取環寬。

陸、 未來展望

為了方便將資料的變數以視覺化的方式呈現，我只考慮了三維的函數進行觀察，若能夠一次放入更多的維度像是死亡、受傷人數等因素，分析結果應該會更豐富，但就需要更加複雜的工具與理論來處理這個問題。

在尋找核密度估計最佳化的過程中找到了許多方法與套件，可惜有許多套件因版本過舊無法使用裡面的函數，若能再加上一些內容(比較 MISE、KDE prediction 等)能使這份報告更加完整。

柒、 資料來源

政府資料開放平台>歷史交通事故資料
110 年度 A1 類交通事故資料
110 年度 A2 類交通事故資料(110 年 1 月- 6 月)
110 年度 A2 類交通事故資料(110 年 7 月-12 月)
https://data.gov.tw/dataset/12197

捌、 參考資料

參考書籍與論文

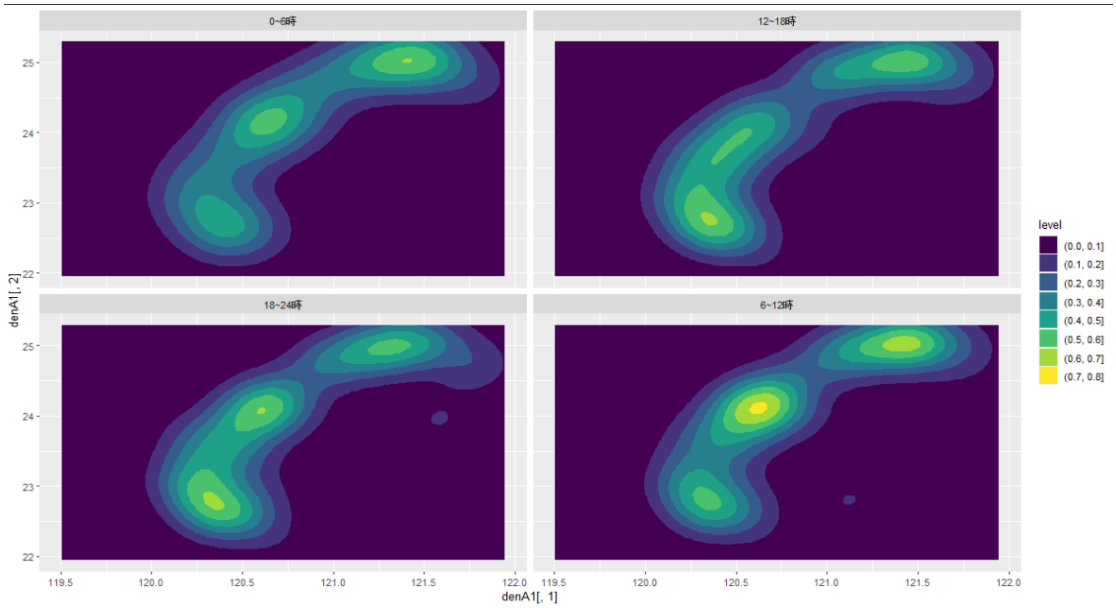
1. Scott, David W., and George R. Terrell. "Biased and unbiased cross-validation in density estimation." *Journal of the american Statistical association* 82.400 (1987): 1131-1146.
2. Sheather, Simon J. "Density estimation." Statistical science (2004): 588-597.
3. Silverman, Bernard W. *Density estimation for statistics and data analysis*. Routledge, 2018.

網址

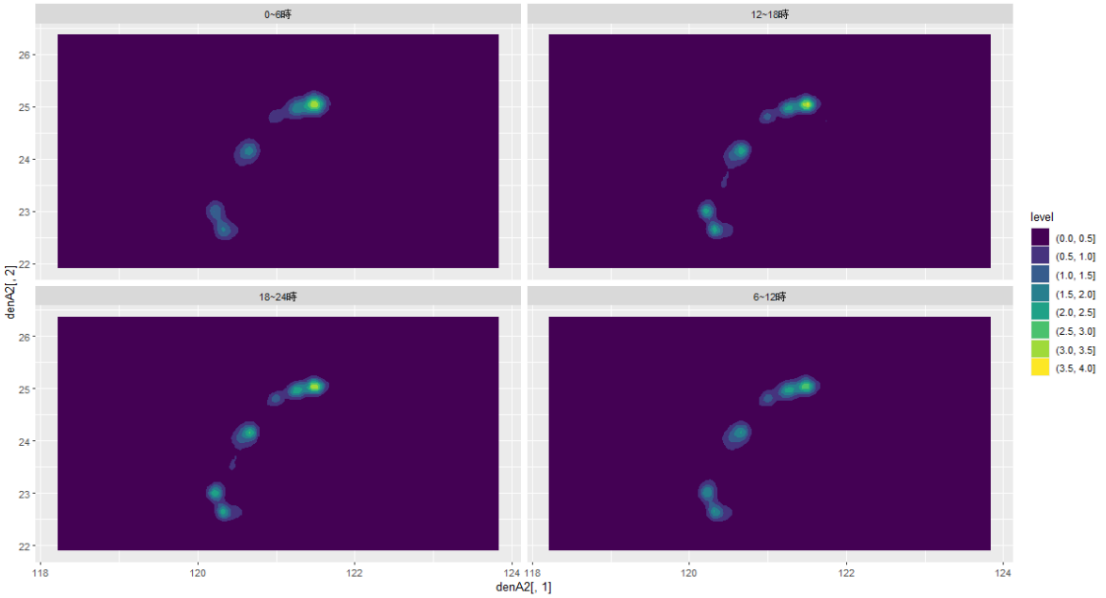
1. https://bookdown.org/egarpor/NP-UC3M/kde-ii-mult.html
2. https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/bcv.html
3. https://scholarship.rice.edu/bitstream/handle/1911/101613/TR87-02.pdf?sequence=1
4. https://stats.stackexchange.com/questions/179043/kernel-density-estimation-bandwidth-rule-of-thumb-2-575-factor
5. https://rdrr.io/cran/misc3d/man/kde3d.html
6. https://stackoverflow.com/questions/60001481/create-contour-in-a-3d-kernel-density-and-find-which-points-are-within-that-co
7. https://cran.r-project.org/web/packages/spNetwork/vignettes/NKDE.html

玖、 附錄

一. 四個時段(0~6、7~12、13~18、19~24 時)台灣交通事故密度估計圖



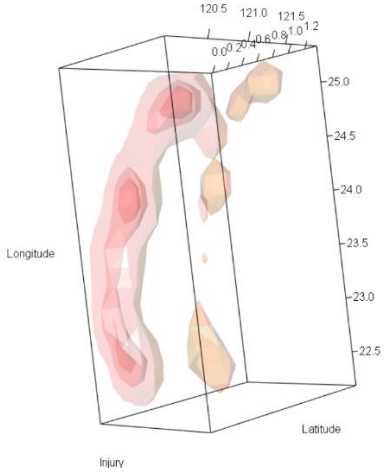
A1 四個時段之密度估計圖



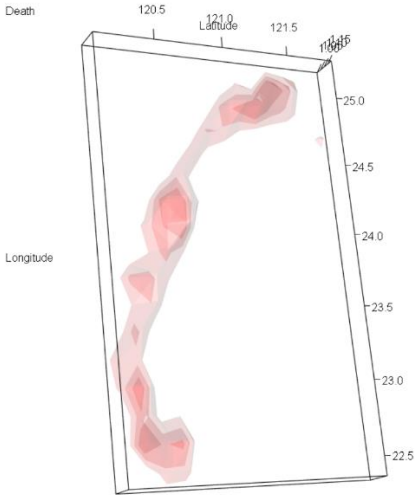
A2 四個時段之密度估計圖

二. A1 及 A2 類經緯度與死傷人數密度估計

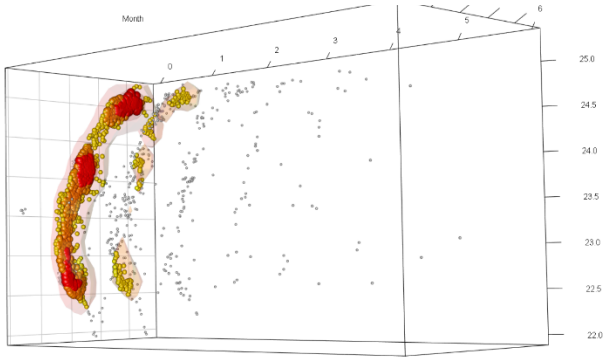
下圖皆以 Gauss kernel function 及環寬以 Rule of thumb 呈現。



A1 事故受傷密度估計圖



A1 事故死亡密度估計圖



A2 事故受傷密度估計圖

```
> summary(dataA1$死亡)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    1.00    1.00    1.03    1.00    6.00
> summary(dataA1$受傷)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.4164  1.0000  7.0000
```

A1 事故死亡、受傷摘要統計

```
> summary(dataA1$受傷)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.4164  1.0000  7.0000
> summary(dataA2$死亡)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0       0       0       0       0       0
```

A2 事故死亡、受傷摘要統計