UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Revista de Topografía Azimut

http://revistas.udistrital.edu.co/ojs/index.php/azimut

Revista de topografía Azimut

Artículo de investigación

Análisis en los residuos de redes altimétricas mínimamente condicionadas

Analysis on the residuals of minimally conditioned altimetric networks

Claudio E. Justo y María Beatriz Pintarelli[(1)]

Para citar este artículo: Justo, C.E. y Pintarelli, M.B. (2021). Análisis en los residuos de redes altimétricas mínimamente condicionadas. *Revista de Topografía Azimut*, Vol. 11 Núm. 1. PP.PP

Fecha de recepción: 21 de marzo de 2020 - Fecha de aceptación: 9 de diciembre de 2020.

Resumen

El análisis de residuos es una parte fundamental del estudio individual de las observaciones en relación con el modelo. Es sabido que la escasez de grados de libertad de las redes altimétricas dificulta esta instancia de su ajuste. Se busca mostrar cómo, con herramientas estadísticas de uso común, es posible realizar una comprobación de residuos a partir de ciertas condiciones. Tales herramientas consisten en estadísticos, histogramas y ploteos de residuos de los cuales valernos para la toma de decisiones. Estas decisiones determinarán la permanencia o no de la observación en el ajuste final. Para obtener observaciones se midió una red existente varias veces, mediante el método de nivelación compuesta con varios

[1] Profesores de la Facultad de Ingeniería. Universidad Nacional de La Plata, Argentina.

niveles automáticos similares. Quedó demostrado que el uso de observaciones duplicadas posibilita construir histogramas y ploteos más cercanos a las hipótesis previas, así como contar con estadísticos capaces de responder estrictamente a una distribución de probabilidades determinada. Para los cálculos se empleó el paquete MASS del software libre R en su versión 3.5.0

Palabras clave: redes altimétricas, método de nivelación compuesta, modelo de ajuste.

Abstract

The analysis of residues is a fundamental part in the individual study of observations in relation to the model. It is known that the scarcity of degrees of freedom of the altimetric networks hinders this instance of its adjustment. The purpose of this work is to show how with common statistical tools it is possible to perform a residue check from certain conditions. Such tools consist of statistics, histograms and plot of residuals that we use to make decisions. These decisions will determine the permanence or not of the observation in the final adjustment. To obtain observations an existing network was measured several times using the compound leveling method with several similar automatic levels. It has been shown that the use of duplicate observations makes it possible to construct histograms and plots closer to the previous hypotheses as well as having statistics capable of responding strictly to a given probability distribution. For the resolution of the calculations, the MASS package of the free software R in version 3.5.0 was used.

Keywords: altimetry networks, compound leveling method, adjustment model.

Introducción

En este trabajo se aplicaron algunas de las técnicas de evaluación de residuos desarrolladas en textos sobre *regresión lineal múltiple*. Aunque mayormente en los trabajos topográficos los modelos funcionales no están en cuestión y se pone bajo estudio a la observación, se optó por este tipo de bibliografía dada la gran variedad de enfoques que ofrece a la hora de estudiar el vector de residuos. Además del análisis de residuos por medios gráficos e inferenciales, se buscó medir la influencia de las observaciones atípicas en los parámetros estimados, cotas en nuestro caso, por medio de la *distancia de Cook*. Los desniveles fueron tomados de sucesivos relevamientos de una red altimétrica sita en el predio de la Facultad de Ingeniería de La Plata (figura 1). Con cada relevamiento se realizó un ajuste por *mínimos cuadrados ponderados* con condicionamiento mínimo en el mismo punto fijo. Para la resolución de los cálculos se empleó el paquete MASS del software libre R en su versión 3.5.0

Desarrollo teórico y metodológico

Modelo de ajuste

El modelo de ajuste por *mínimos cuadrados ponderados* empleado en este trabajo queda expresado de la siguiente manera:

$$\boldsymbol{y = A.X + C + \varepsilon} \qquad (1)$$

Donde el vector de desniveles *y* se modela linealmente mediante las cotas o parámetros *X*, un vector de constantes *C* y un vector *ε* de ruido aleatorio. El estudio de este vector *ε*, luego de ser estimado, es el motivo de este trabajo. Este vector tiene por hipótesis distribución normal con *E(ε)=0* mientras que su varianza está expresada por la matriz de varianzas covarianzas *V* ≡ *Var(ε)*:

$$\boldsymbol{Var(\varepsilon) = \sigma^2\, Q}$$

$$Var(\Delta H_{AB}) = \sigma_{km}{}^2\, Q_{AB} = V$$

El valor $\sigma_{km}{}^2$ es la varianza *a priori* que resultaría de nivelar un kilómetro en forma simple. La matriz *Q* es la matriz de cofactores es diagonal debido a que las observaciones son estadísticamente independientes. En la diagonal se encuentran la cantidad de kilómetros empleados para la determinación de cada desnivel. Nuestra matriz de pesos *P* provendrá de la inversa de *V*. El vector *C* es el que permite introducir el *datum* mediante un condicionamiento mínimo. Este tipo de condicionamiento implica que solo las observaciones determinarán la calidad del ajuste con prescindencia del marco de referencia. La sobredeterminación del sistema de ecuaciones lineales

$$\boldsymbol{y = A.X + C + \varepsilon}$$

es salvada mediante la aplicación del criterio de *mínimos cuadrados ponderados*, donde el sistema de *ecuaciones normales*

$$\boldsymbol{A^t.P.A.\tilde{X} = A'.P.(y + C)}$$

permite obtener una única solución $\hat{\boldsymbol{X}}$ para el vector de cotas $\hat{\boldsymbol{X}}$. Dado que las observaciones pertenecen a una distribución Normal el estimador de *mínimos cuadrados* es también un estimador de *máxima verosimilitud*. A partir de este obtenemos el vector de residuos estimados *e*:

$$\boldsymbol{e} = \boldsymbol{A}.\hat{\boldsymbol{X}} - (\boldsymbol{y} + \boldsymbol{C})$$

Estudios posibles sobre los residuos estimados en el ajuste

El hecho de que luego del ajuste los residuos tendrán diferente varianza, aunque *a priori*, sean las observaciones del mismo peso o precisión (Montgomery, Peck y Vining, 2007), implica que debamos escalarlos. Esto debe hacerse para hacerlos comparables en un mismo estudio ya sea *gráfico* o mediante la construcción de *estadísticos*.

Esto puede hacerse estandarizando o estudentizando los residuos. Lo primero se logra mediante la expresión,

$$\boldsymbol{d_i} = \frac{\boldsymbol{e_i}}{\sqrt{\boldsymbol{s^2}}}$$

Donde s^2 es la varianza del ajuste. Estos residuos tendrán media cero (0) y varianza aproximadamente unitaria (Montgomery, Peck y Vining, 2007).

La estudentización se realiza mediante la expresión,

$$\boldsymbol{r_i} = \frac{\boldsymbol{e_i}}{\boldsymbol{s_{e_i}}} = \frac{\boldsymbol{e_i}}{\sqrt{\boldsymbol{s^2}.\boldsymbol{q_{e_i}}}}$$

Donde S_{e_i} es la varianza de cada residuo estimado, ya para lo cual deberemos contar con la matriz de varianzas covarianzas (Ghilani y Wolf, 2006) de los residuos, la cual se expresa:

$$S_e^2 = S^2.(P^{-1} - A.(A^t.P.A)^{-1}.A^t) = S^2.Q_e$$

Cuando el modelo es correcto estos residuos r_i tienen varianza constante $\sigma_{r_i}^2 = 1$ y para gran cantidad de datos no habrá mucha diferencia entre los residuos estandarizados y estudentizados (Montgomery, Peck y Vining, 2007).

Estudios gráficos

Los residuos así escalados podrán estudiarse en histogramas o en *Q-Q plots*, ya que pertenecerán a una misma población. Hay que recordar al tomar la decisión de recurrir a este tipo de herramientas gráficas que se recomienda su empleo cuando el número de observaciones es superior a 20 (Montgomery, Peck y Vining, 2007).

Estudios mediante estadísticos

Los r_i serán estadísticos con una distribución *tipo t* [2] con *n-r* grados de libertad. Estrictamente no pertenecen a una distribución *t* por no ser independientes el numerador del denominador (Montgomery, Peck y Vining, 2007). Este tipo de escalamiento donde se incluye a la observación sospechada de atípica se denomina *interno*.

Para resolver la falta de independencia entre numerador y denominador, se estima $S^2_{(i)}$ prescindiendo de la observación Δh_i considerada atípica en función de su residuo r_i. Esto

independiza numerador y denominador, y evita un escalamiento interno (Montgomery, Peck y Vining, 2007).

Se obtiene un estadístico con distribución *t de Student* con *n-p-1* grados de libertad. La estimación de $S^2_{(i)}$ puede ser realizada con la expresión

$$S^2_{(i)} = \frac{(n-p).S^2 - \frac{e_i^2}{(1-h_{ii})}}{n-p-1}$$

Recordando que $(1-h_i)$ es q_{e_i}.

Esta expresión evita hacer los ajustes nuevamente. Así el estadístico de estudio será

$$R_i = \frac{e_i}{\sqrt{S^2_{(i)} q_{e_i}}}$$

Esto último solo será posible cuando contemos con observaciones duplicadas.

Una vez que contemos con estos estadísticos podremos emplear la distribución T Student para decidir si su valor es significativamente distinto de cero ya sea para un α=0.05 o 0.01. Si el módulo del residuo supera el valor de corte establecido para α y para *n-p* se tendrá un elemento de juicio para su estudio en particular o reemplazo por una nueva observación.

Influencia de las observaciones

Cuando es encontrada una observación atípica es recomendable obtener una medida de su influencia o por lo menos verificar su influencia en los parámetros mediante la

comparación directa de las soluciones. La distancia de Cook (Montgomery, Peck y Vining, 2007) es un estadístico que mide si la distancia cartesiana, elevada al cuadrado, entre dos soluciones es significativa. Estas soluciones se diferencian en que una de ellas no incluye al punto sospechado. Esta distancia puede expresarse en función de los parámetros de la siguiente manera:

$$D_i = \frac{(X_{(i)} - X)^t . A^t . P . A . (X_{(i)} - X)}{p . S^2}$$

En la práctica, suele emplearse el siguiente estadístico (Montgomery, Peck y Vining, 2007).

$$D_i = \frac{r_i^2}{p} \left(\frac{s_{l_i}^2}{s_{e_i}^2} \right)$$

Donde r_i es el residuo estudentizado y p la cantidad de parámetros del ajuste. D_i no es un estadístico F; sin embargo, usar el valor de corte $F_{0.5,p,n-p} \approx 1$ funciona bien en la práctica (Montgomery, Peck y Vining, 2007). Es decir que se mide la significancia de una observación presuntamente influyente. Lo usual es poner en duda la observación cuya $D_i \geq 1$.

Aplicación a un caso real

Los conceptos presentados fueron aplicados con datos de diversos relevamientos realizados en una red altimétrica situada en el campus de la Facultad de Ingeniería de La Plata. En la

vista aérea se aprecian las ubicaciones de las marcas físicas vinculadas mediante diferencia de alturas. Como *datum*, se adoptó el punto AGRIM_VIEJA (AV).

Figura 1. Vista de la Red Altimétrica de la Facultad de Ingeniería de La Plata

Las observaciones fueron tomadas, por estudiantes de la carrera Ingeniero Agrimensor, con niveles automáticos de similares características y durante distintos cursos del dictado de la asignatura Cálculo de Compensación.

La condición necesaria para la aplicación de test estadísticos basados en las distribuciones *chi* cuadrado, *t de Student* y Fisher es la pertenencia de las observaciones a una distribución normal. Esto no puede hacerse para cada desnivel en forma individual en vista de la poca

cantidad de datos, por cada desnivel; pero sí es posible verificar tal circunstancia en los residuos del ajuste de una red con muchos grados de libertad. Luego de ajustar la totalidad de las 58 observaciones en total, con 51 grados de libertad se realizó una bondad de ajuste por *chi* cuadrado de los residuos estandarizados encontrando un *p* valor de 0,4277, por lo que no se encontraron elementos para rechazar la normalidad de los datos.

Aplicación a una de las redes

En la tabla 1 se detallan las observaciones realizadas en la red de la imagen aérea, donde estas se ajustan junto con los residuos brutos *e*, los residuos estudentizados *r* y los residuos *r de Student*. En la última columna se ven los valores de la distancia de Cook.

Tabla 1. Valores resultantes del ajuste mínimamente condicionado

Tramo medido	DH bruto	DH ajustado	*e*	r	R	Dist, Cook
AV - AN	+2,046	+2,0460423	-0,0004231	-0,30553	-0,2946033	0,0085646
Q1 - AN	+1,110	+1,1100164	-0,0000164	-0,010755	-0,0103331	0,0000084
D – Q1	+0,112	+0,1105502	0,0014498	1,3849766	1,4412095	0,1822203
Q2 – D	+0,381	+0,3764464	0,0045536	3,2334299	7,0212768	0,7538931
P - H	-1,615	-1,6148721	-0,0001279	-0,0727518	-0,0699119	0,0003866

C – Q2	-0,920	-0,9212172	0,0012172	1,0037532	1,0040679	0,0849196
AN - H	+0,166	+0,1661877	-0,0001877	-0,1491127	-0,1433856	0,0016492
C-Q1	-0,435	-0,4342206	-0,0007794	-0,601502	-0,5861182	0,0233012
P - AV	+0,267	+265363	0,0016367	0,9227019	0,9170405	0,0615705
C - H	+0,842	+0,8419835	0,0000165	0,010763	0,0103408	0,0000084
AN - AV	-2,046	-2,0460423	0,0004231	0,30553	0,2946033	0,0085646
AN – Q1	-1,110	-1,1100164	0,0000164	0,010755	0,0103331	0,0000084
Q1 – D	-0,111	-0,1105502	-0,0004498	-0,429699	-0,4158049	0,0175404
D – Q2	-0,375	-0,3764464	0,0014464	1,0270777	1,0294347	0,0760658
H - P	+1,616	+1,6148721	0,0011279	0,641391	0,6262164	0,0300517
Q2 - C	+0,920	+0,9212172	-0,0012172	-1,0037532	-1,0040679	0,0849196
H-AN	-0,167	-0,1661877	-0,0008123	-0,6451983	-0,6300562	0,0308775
Q1-C	+0,436	+0,4342206	0,0017794	1,3732679	1,4269477	0,1214548
AV-P	-0,265	-0,26535	0,0003633	0,2048246	0,1971074	0,003034
H - C	-0,842	-0,8419835	-0,0000165	-0,010763	-0,0103408	0,0000084

Esta red resultó con una varianza del ajuste de $\boldsymbol{S^2} = 0.00002304\ \text{m}^2$ con $\boldsymbol{n-p=13}$ grados de libertad. La varianza $\sigma_{Km}^2 = 0{,}000025\ \text{m}^2$ fue satisfactoriamente testeada bilateralmente mediante *chi* cuadrado para $\boldsymbol{\alpha = 0.05}$. Primero la magnitud de los residuos estudentizados *r* se analizó para $\boldsymbol{\alpha = 0.05}$ con un $\boldsymbol{t_{\frac{\alpha}{2}} = 2.16}$ (Walpole, Myers y Myers, 1998) y la de los residuos *R*, al tener $\boldsymbol{n-p-1=12}$ grados de libertad, para un

$t_{\frac{\alpha}{2}} = 2.18$ (Walpole, Myers y Myers, 1998). Se encontró que el valor *i=4* alcanzó el valor $r = 3.23$ superior al $t_{\frac{\alpha}{2}} = 2.16$ (Walpole, Myers y Myers, 1998), considerándola atípica. Cuando se realiza el escalamiento externo del residuo bruto para obtener el residuo R con 13-1 grados de libertad se obtiene un valor de 7,02. En la figura 2 tenemos tanto el *Q-Q plot* de los residuos *r* como el gráfico de los residuos *r* y *R* ,en ambos casos contra los valores ajustados de los desniveles. Finalmente, el gráfico de las Distancias de Cook en función de los mismos desniveles ajustados se ve en la figura 3.

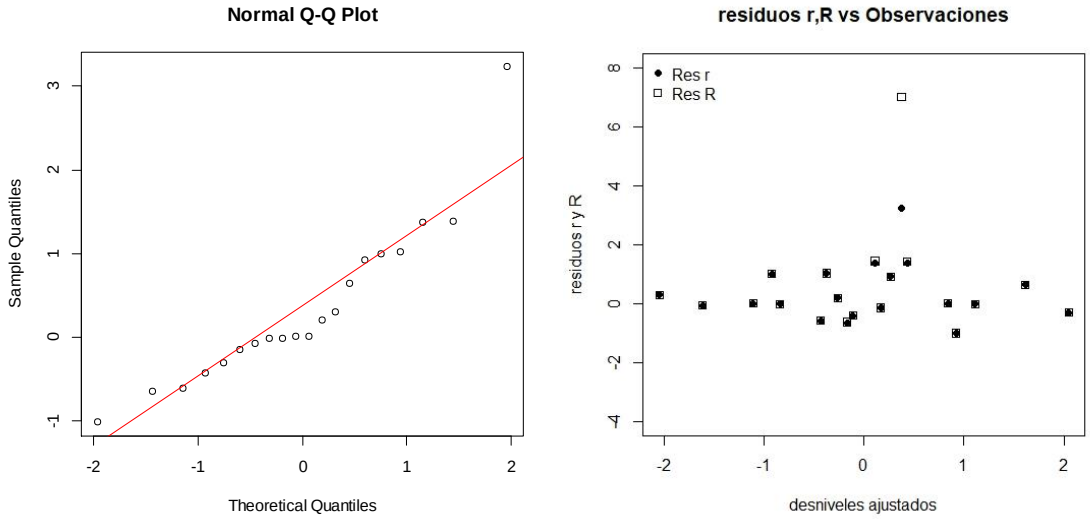

Figura 2.Fuente:Q-Q plot de residuos r **Residuos r y R vs. Desniveles ajustados**

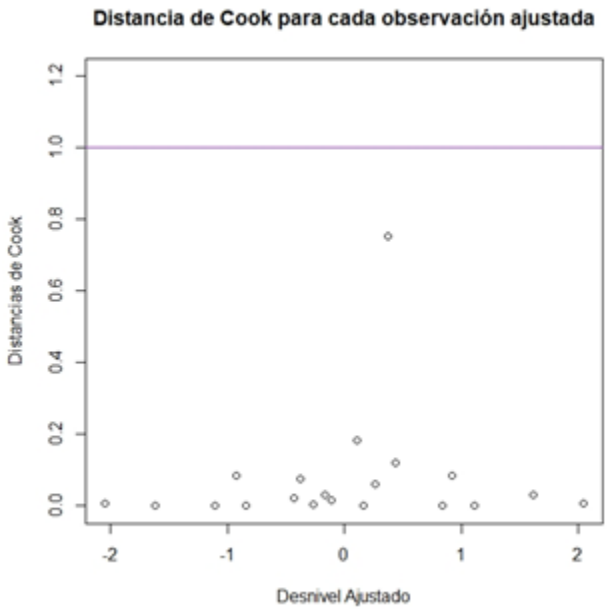

Figura 3.Fuente: Distancias de Cook vs. Desniveles Ajustados

Los cálculos se realizaron empleando el paquete MASS del software R en su versión 3.5.0 . con la función: lm (formula = y ~x - 1, weights = w) . Siendo *y* el vector de las observaciones y *x* el vector de parámetros a estimar, en nuestro caso las cotas de las ménsulas. El vector w contiene los pesos de las observaciones y el valor -1 indica que en la regresión no existe término independiente.

Hallazgos y conclusiones

El empleo de técnicas gráficas como el *Q-Q plot* y el ploteo de residuos contra valores ajustados muestra rápidamente el cumplimiento, o no, de hipótesis, como adecuación al modelo y homogeneidad de varianzas, así como también la existencia de valores atípicos. La decisión de rechazar una observación en función de la construcción de estadísticos con distribución *t de Student* debe adecuarse a la cantidad de datos. Con redes cuyos desniveles

fueron medidos solamente una vez los residuos brutos solo pueden estudentizarse de forma interna, con los residuos *r* ya que no puede excluirse ninguna de las observaciones sin perder parte de la red. La estudentización externa, mediante residuos *R* y aplicable solo a redes con observaciones dobles, mostró cómo los residuos reaccionan en caso de tener una varianza estimada sin el aporte del dato atípico. La distancia de Cook no mostró ser sensible al valor atípico. Cabe recordar que las circunstancias que generen el apartamiento del valor esperado se distribuyen a toda la red. La distancia de Cook mide la influencia en todos los parámetros estimados. Vale considerar que existen estadísticos que permiten estudiar la influencia de una observación atípica específicamente sobre cada uno de los parámetros considerados. Como recomendación general en el empleo de las herramientas estadísticas, estas deben emplearse en forma complementaria para confirmar los hallazgos mediante las diferentes técnicas disponibles.

Referencias

Ghilani, Ch. y Wolf, P. (2006). *Adjustment computations*. 4a. ed. Ciudad New York: Wiley & Sons.

Montgomery, D., Peck, E. y Vining, G. (2007). *Introducción al análisis de regresión*. 3a. Ciudad de México: Ed. Patria.

Walpole, R., Myers, R. y Myers, S. (1998). *Probabilidad y estadística para ingenieros*. 6a. ed. México: Pearson Educación.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Revista de Topografía Azimut

http://revistas.udistrital.edu.co/ojs/index.php/azimut

Revista de topografía
Azimut