ANALELE ŞTIINŢIFICE ALE UNIVERSITĂŢII „ALEXANDRU IOAN CUZA” DIN IAŞI
Tomul LVI **Ştiinţe Economice** **2009**

USING INFLUENCE FUNCTION MATRIX AS OUTLIER DETECTING TOOL BASED ON POOLED SERIAL CORRELATION COEFFICIENTS

S.R.T. MOENG, D.K. SHANGODOYIN, P.M. KGOSI*

Abstract

In this paper, we incorporated autocorrelation function (ACF), partial autocorrelation function (PACF) and inverse auto-correlation function (IACF) into the influence function as a graphical tool for detecting outliers. Depending on the number of positive and negative values of the influence function based on critical values obtained for different lags an observation is identify as outlier. Both the simulated data and Botswana meat sales data confirms the efficacy of using the pooled correlation coefficients in influence function matrix as outlier detection device.

Keywords: Outliers, Critical values, ACF, PACF, IACF and influence function
JEL classification: C01, C10

1. Introduction

Structural changes occur frequently in Time Series data which can jeopardize the use of some correlation functions as model identification and estimation tools. Shangodoyin (2000) has shown that the structural changes have some implications on the Box and Jenkins methods of model fitting. Model estimation and identification techniques in time series made use of autocorrelation function coefficients(ACF) for which its estimators are not totally free of bias; the expected value of two well known estimators of ACF have been shown(Wei(2007)) as $E(\hat{\gamma}) \approx \gamma_k - \frac{k}{n}\gamma_k - \left(\frac{n-k}{n}\right)Var(\bar{X})$ and $E\left(\hat{\hat{\gamma}}\right) \approx \gamma_k - Var(\bar{X})$, from these expectations it clear that the estimators are biased except for large values of n. This challenge has made researchers to exploit the use of partial autocorrelation function coefficients (PACF) and inverse autocorrelation coefficients in modeling; for instance, Shangodoyin (1998) has extended the influence function to partial autocorrelation function and concluded that it surpasses the autocorrelation function in detecting significant outliers in series, and Olewuezi (2007) has extended the influence function to inverse autocorrelation coefficients but did not compare it performance to Partial auto- correlation coefficients. In this paper, we investigate the detection of outliers in time series data by considering the influence function for ACF, PACF and IACF at specified positive lag (k).

* S.R.T. Moeng, D.K. Shangodoyin, P.M. Kgosi, University of Botswana.

2. The Influence Function for correlation Coefficients

Influence function measures the effect of a small change of data distribution on the value of a statistic or an estimate of interest. Following Chernick et al. (1982) and Shangodoyin (1998), let $\rho(k)$ be the auto-correlation function at lag k, for a stationary series $\{X_t\}_{t=1}^{N}$ with $E(X_t)=\mu$ and $Var(X_t)=\sigma^2$. Without loss of generality on $\rho(k)$, suppose we assume $Z_t=\dfrac{X_t-\mu_t}{\sigma}$ for every value of t. Then the influence function

$$I\left[F,T(F),X\right]=\lim_{\varepsilon\to 0}\frac{T(1-\varepsilon)F+\varepsilon\delta_X-T(F)}{\varepsilon} \tag{1}$$

where $T(F)$ is the estimator under consideration and the empirical distribution F becomes $I\left(H,\rho(k),Z\right)$. That is, $I\left[F,\rho(k),X\right]=I\left[H,\rho(k),Z\right]$ and H is the distribution for (Z_t,Z_{t+k}).

Let $U_{i,k,1}=\dfrac{1}{2}\left[\dfrac{Z_i+Z_{i+k}}{\sqrt{1+\rho(k)}}+\dfrac{Z_i-Z_{i+k}}{\sqrt{1-\rho(k)}}\right]$ and $U_{i,k,2}=\dfrac{1}{2}\left[\dfrac{Z_i+Z_{i+k}}{\sqrt{1+\rho(k)}}-\dfrac{Z_i-Z_{i+k}}{\sqrt{1-\rho(k)}}\right]$. Then we have $\left[1-\rho^2(k)\right]^2U_{i,k,1}U_{i,k,2}=z_iz_{i+k}-\dfrac{1}{2}\rho(k)\left(z_i^2+z_{i+k}^2\right)$. Therefore, the influence function based on $\rho(k)$ (see Chernick et al. (1982)) is

$$I\left[H_1\,\rho(k),(z_i,z_{i+k})\right]=\left[1-\rho^2(k)\right]U_{i,k,1}U_{i,k,2} \tag{2}$$

Because of the similarity between ρ_k (ACF) and α_{kk} (PACF) as model identification and estimation tools, Shangodoyin (1998) has extended (2) to α_{kk} as

$$I\left[H,\alpha(k),(z_i,z_{i+k})\right]=\left[1-\alpha^2(k)\right]U_{i,k,1}U_{i,k,2} \tag{3}$$

Cleveland (1972) has introduced the concept of inverse autocorrelation function (IACF) given as

$$\rho_i(k)=\frac{\gamma_*(k)}{\gamma_*(0)} \qquad \gamma_i(k)=\int_{-\pi}^{\pi}e^{ikw}f_*(w)\,dw \tag{4}$$

where $\gamma_*(k)=\gamma_*(-k)$ for all $k=0,1,2,\ldots$ and $f_*(w)=\left\{f(w)\right\}^{-1}$ is the inverse spectral density function. Cleveland (1972) has shown that IACF contains all the information in the PACF including the information of possible presence of zero at some lags. Olewuezi and Shangodoyin (2005) relying on the spectrum in (4) have given the relationship between IACF and ACF as

$$\gamma_*(k)=\begin{cases}\dfrac{2\pi}{\gamma_0}, & k=o\\ 4\pi\sum_{j=1}^{k}\left(\dfrac{1}{J}\right)\left(\dfrac{1}{\gamma_J}\right)\sin 2j\pi, & k\neq 0\end{cases} \tag{5}$$

Exploiting the duality between IACF and ACF as given in (4) and (5), we can extend equation (2) to IACF to get the influence function as

$$I\left[H,\rho_i(k),(z_i,z_{i+k})\right]=\left[1-\rho_i^2(k)\right]U'_{i,k,1}U'_{i,k,2} \tag{6}$$

Where $U'_{i,k,1}=\frac{1}{2}\left[\frac{z_i+z_{i+k}}{\sqrt{1+\rho_i(k)}}+\frac{z_i-z_{i+k}}{\sqrt{1-\rho_i(k)}}\right]$ and $U'_{i,k,2}=\frac{1}{2}\left[\frac{z_i+z_{i+k}}{\sqrt{1+\rho_i(k)}}-\frac{z_i-z_{i+k}}{\sqrt{1-\rho_i(k)}}\right]$.

Assuming a stationary Gaussian process, for which μ, σ, $\rho(k)$, $\alpha(k)$ and $\rho_i(k)$ are known, then $U'_{i,k,1}$ and $U'_{i,k,2}$ are observations from independent standard normal distributions, the quantity $I\left[H,\rho_i(k),(z_i,z_{i+k})\right]$ has the distribution of a constant multiplied by a product of standard normal random variables.

3. Outlier Detection Scheme

To investigate the effect of outliers on the dataset, we made use of the estimates in equations (2), (4) and (6). . To do this, we construct the influence function matrix for all the autocorrelation functions discussed in section 2 and set the critical value based on two standard errors of the mean for the computed influence function values.

Given an $n\times k$ matrix (where n is the number of observations and $k\leq\frac{n}{4}$ is a fixed lag) with the $(j,k)^{th}$ entry given by $I\left[H,\rho(k),(U_j,U_{j+k})\right]$ or $I\left[H,\rho_i(k),(U'_j,U'_{j+k})\right]$ with the quantity U_j and $U_j{}'$ being the j^{th} standardized observation. Although Chernick et al. (1982) have pointed out that observation U_t influences several lagged autocorrelation estimates and appears in the computation of every element in the i^{th} row and also in the diagonal elements of the preceding rows beginning in column 1 of row $i-1$ and preceding to the right this can only apply to negative lag (k). A review of the paper by Chernick et al (1982) with the findings of Shangodoyin (1998) and Olewuezi (2007) reveals that by computing the influence function matrix based on the pool effect of ACF, PACF and IACF in a single matrix will complement one another and identify outliers for any given sample size ; these functions will complement one another in identifying the outliers in the dataset when used in this way and will also serve as a palliative measure to alleviate the challenges of using these functions as discussed Wei (2007).

An outlier will often have a very large positive (+) (if the influence function estimate is positive and is significant based on the specified critical region) or negative (-) (if the influence function is negative and is significant based on the specified critical region), while the other observations are left blank.

The matrix will then appear with patterns '+'and '-', the rows of the combine influence function of ACF, PACF and IACF with large values of "+" and "−" signified that the observations at those points are outliers. We shall demonstrate the use of these methodology using both simulated data and Botswana's sales of meat (local) data in the next section.

4. Data Analysis using simulated data and Botswana Meat Data

(I) **Simulated Data**: The NRAN (Normal Random number generated) module of SYSTAT was used to simulate the series with contamination (induced outliers) at time point T= 17, 42 and 96. The contaminated series has standard deviation 82.66 and mean 66.855, where as uncontaminated series has standard deviation 46.422 and mean 57.072. The coefficient of variation (CV) for contaminated series is 123% and is very high compare with CV of 81% for uncontaminated series. The values of ACF, PACF and IACF at lag one for contaminated series are respectively -0.48244, -0.48244 and 0.64601 compared with the values of ACF, PACF and IACF for contaminated series which are -0.50595, -0.50595 and 0.91824; the values were blown up in the latter situation.

The values of ACF, PACF and IACF from ARMA (0,1,1) model, for both contaminated and uncontaminated series, are shown in Table 1. The values of ACF, PACF and IACF at lag one for contaminated series are respectively -0.48258, -0.48258 and 0.77792 compared with the values of ACF, PACF and IACF for contaminated series which are --0.49879, -0.49879 and 0.95281; the values were blown up in the latter situation.

Table 1: Correlations for the contaminated and uncontaminated simulated series at different lags.

Lag	ACF		PACF		IACF	
	a	b	a	b	a	b
1	-0.48258	-0.49879	-0.48258	-0.49879	0.77792	0.95281
2	-0.00004	-0.00032	-0.30364	-0.33162	0.63765	0.90622
3	0.00178	-0.00941	-0.20790	-0.26179	0.55354	0.86004
4	-0.00146	0.01872	-0.15095	-0.19121	0.50016	0.81419

a -Simulated uncontaminated series b -contaminated series

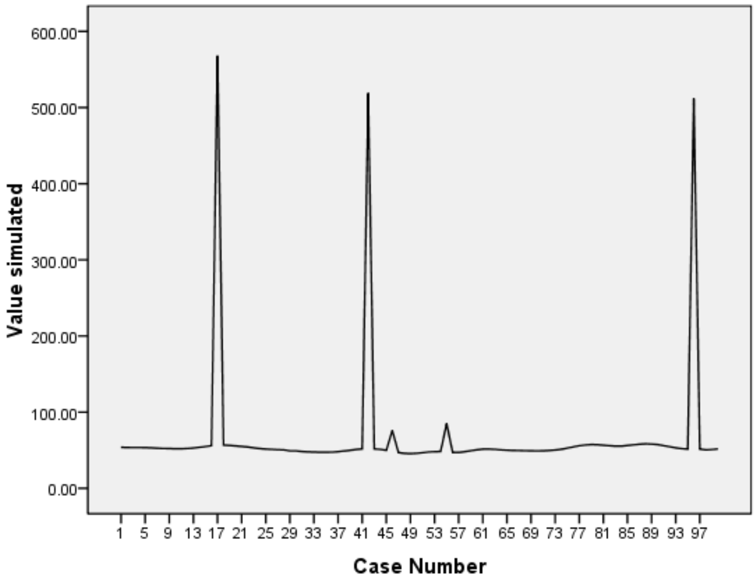

Figure 1: Simulated Data With Induced Outliers

From these summary statistics and Figure 1, we can deduce that the effect of few outliers on the estimates of autocorrelation, partial autocorrelation and inverse autocorrelation could jeopardize their function as model identification tools. The influence function matrix was used to identify observations with large influence based on the critical value of two standard errors of the mean for the computed values of influence function. The critical values for different functions at specified positive lags are displayed in Table 2 below.

Table 2: Critical values for the autocorrelation functions for different lag periods, for the simulated data.

Lag	Critical values		
K	ACF Mean $\pm$ SE	PACF Mean $\pm$ SE	IACF Mean $\pm$ SE
1	0.7851 ± 0.28299	0.7851 ± 0.2899	-0.00858 ± 0.003994
2	0.0292 ± 0.02418	0.2373 ± 0.07510	-0.9444 ± 0.37717
3	-0.0259 ± 0.20358	0.3458 ± 0.026456	-0.9105 ± 0.37822
4	-0.0323 ± 0.02582	0.2233 ± 0.26825	-0.8591 ± 0.36353

Based on the critical values displayed above, we derived the influence function matrix in Appendix I, in this appendix, we found that at time T=17,42 and 96 the influence matrix reveals that the observations influence several lagged ACF, PACF and IACF (for this it is not pronounced in lag one) estimates and it appears in the computation of every element in the 17th, 42nd and 96th rows and also in the diagonal element of the subsequent rows beginning in column 1 of row $i+1$ and proceeding to the right downwards. This is an inverse of what Chernick et (1982), Shangodoyin (1998) and Olewuezi(2007) obtained due to the sign of the lag (k) used in the computation of the values of $U_{i,k,1}$ and $U_{i,k,2}$ discussed in section 2.This is not a challenge as the ACF, PACF and IACF are even functions(see Chatfield (2003) and Shangodoyin & Ojo (2002)). It is interesting to note that ACF and PACF behave in similar manner and the combine Influence functions matrix of ACF, PACF and IACF has significantly led to the detection of the outliers at period 17, 42 and 96 with no other point identified as outlier.

(II)**Botswana meat sales (local) data:** From the sales data see Appendix II, it appears that observations at time points T= 19, 26, 27 and 28 are outliers compare with the rest of the observations (see Fig 2)

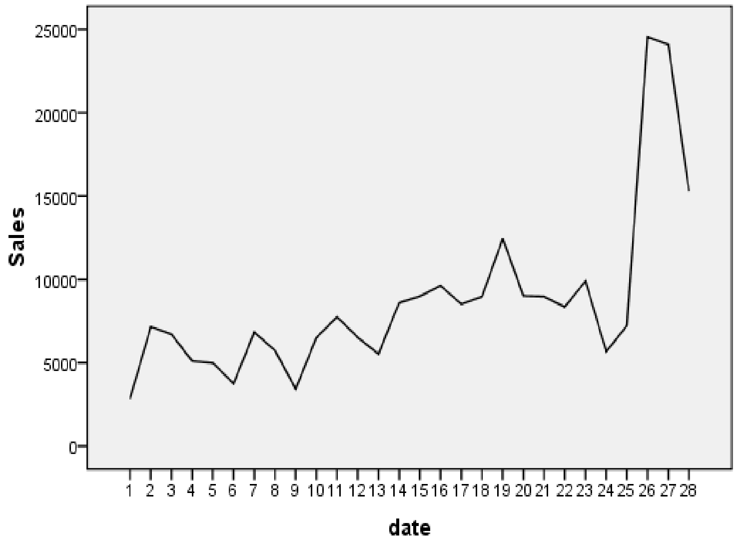

Figure 2: Graph Of Bmc Sales Data

Figure 2 gives room for questioning the observations at these time points, thus we apply the methodology described in section 3 to ascertain the status of the observations at these time points. The critical values for different functions at different lags are displayed in table 3.

Table 3: Critical values for the autocorrelation functions for different lag periods, for the BMC sales data

Lag	Critical values		
K	ACF	PACF	IACF
1	0.6463 ± 0.7638	0.6459 ± 0.7635	0.4239 ± 0.5486
2	1.1066 ± 0.9581	1.118 ± 0.9670	-0.361 ± 0.4880
3	0.774 ± 0.7599	0.8015 ± 0.7706	-0.0329 ± 0.3061
4	0.7808 ± 0.7587	0.8463 ± 0.8210	-0.1011 ± 0.2894

Based on the critical values in table II above we construct the influence function matrix given in Appendix II. It reveals that large positive and negative values are noticed in rows 26, 27 and 28. This indicates that the observations at these points are outliers, one could check on the producer of this data and if found that they are actual recordings then, analyst may wish to determine the actual magnitude of the outliers by using the technique provided in Shangodoyin and Arnab (2005).

5. Conclusions

We have investigated the use of ACF, PACF and IACF in detecting outliers with critical values based on two standard deviations of the influence function estimates; this situation was not captured by Chernick et al (1982), Shangodoyin (1998) and Olewuezi (2007). The advent of statistical software has made it possible for analysts to derive the influence function matrix for ACF, PACF and IACF in a single sweep for different lags, this will help analyst to decide on the status of a suspected outlying observation. It is interesting to note that (the influence detection using) both the ACF and PACF behave in similar manner, while IACF pattern differs. It is worth noting that in this scenario, IACF does not behave in a similarly with PACF contrary to the suggestion by Wei (2007) page 275.

In addition this paper confirms the influence (effect) of the outliers, in a time series data, in the estimation of correlation measures and subsequent statistical estimation based on these measures.

6. References

Chatfield C (2003): The analysis of time series: An introduction. Sixth Edition. Chapman & Hall/CRC

Chernick, M.R, Downing, D.I and Pike, D.H (1982): Detecting outliers in Time series data. Journal of American Statistical Association. Vol. 17, No. 380, pp 743-747

Cleveland W. S (1972): The inverse autocorrelations of a time series and their application. Technometrics, 14, 277-298

Olewuezi, N.P and Shangodoyin, D.K (2005): Contributions of inverse autocorrelation function in transfer function modeling. Far East Journal of theoretical Statistics. Vol. 17(2), pp 227-238.

Olewuezi, N.P (2007): Contributions of influence function using the inverse autocorrelation in detection of outliers. Journal of the Nigerian Association of Mathematical Physics. Vol. 11 pp627-634

Shangodoyin D .K (1998) : The use of correlation function for detecting outliers ad model order determination. Journal of Nigerian Statisticians. Vol. 1 pp22-28

Shangodoyin, D.K (2000): Specification of dynamic distributed lag model connecting two outlier infested time series. Journal of Science Research. Vol. 6 pp 97-103

Shangodoyin, D.K & Ojo, J.F (2002): Elements of Time series analysis. Rasmed Publications. ISBN: 978-8024-32-7

Shangodoyin, D.K & Arnab, R (2005): The computational efficiency of outlier estimates for different generating process. Model assisted Statistics and Applications. Vol. 1 #2, pp69-75

Wei (2007): Time series analysis: Univariate and Multivariate methods (5th edition). Addison Wesley.

APPENDIX I

Table A1: Sign matrix for influence values identifying outliers for simulated data with induced outliers

		ACF				PACF				IACF			
		Lag (k)				Lag (k)				Lag (k)			
simulated	date	1	2	3	4	1	2	3	4	1	2	3	4
53.8	1												
53.6	2	-				-				+			
53.5	3	-	+			-	-			+	+		
53.5	4	-	+			-	-			+	+	+	
53.4	5	-	+			-	-			+	+	+	+
53.1	6	-	+			-	-			+	+	+	+
52.7	7	-	+			-	-			+	+	+	+
52.4	8	-	+			-	-			+	+	+	+
52.2	9	-	+			-	-			+	+	+	+
52	10	-	+			-	-			+	+	+	+
52	11	-	+			-	-			+	+	+	+
52.4	12	-	+			-	-			+	+	+	+
53	13	-	+			-	-			+	+	+	+
54	14	-	+			-	-			+	+	+	+
54.9	15	-	+			-	-			+	+	+	+
56	16	-	+			-	-			+	+	+	+
568	17	+	-	-	-	+	+	+	+	-	-	-	-
56.8	18	+	+			+	-			-	+	+	+
56.4	19	-	-	+	+	-	+	+	+	+	-	+	+
55.7	20	-	+	-		-	-	-		+	+	-	+
55	21	-	+		-	-	-		-	+	+	+	-
54.3	22	-	+			-	-			+	+	+	+
53.2	23	-	+			-	-			+	+	+	+
52.3	24	-	+			-	-			+	+	+	+
51.6	25	-	+			-	-			+	+	+	+
51.2	26	-	+			-	-			+	+	+	+
50.8	27	-	+			-	-			+	+	+	+
50.5	28	-	+			-	-			+	+	+	+
49.2	29	-	+			-	-			+	+	+	+
49.2	30	-	+			-	-			+	+	+	+
48.4	31	-	+			-	-			+	+	+	+
47.9	32	-	+			-	-			+	+	+	+
47.6	33	-	+			-				+	+	+	+
47.5	34	-	+			-				+	+	+	+
47.5	35	-	+			-				+	+	+	+
47.6	36	-	+			-				+	+	+	+
48.1	37	-	+			-				+	+	+	+
49.1	38	-	+			-	-			+	+	+	+
50	39	-	+			-	-			+	+	+	+
51.1	40	-	+			-	-			+	+	+	+
51.8	41	-	+			-	-			+	+	+	+
519	42	+	-	-	-	+	+	+	+	-	-	-	-
51.7	43	+	+			+	-			-	+	+	+
51.2	44	-	-	+	+	-	+	+	+	+	-	+	+
50	45	-	+	-		-	-	-		+	+	-	+
76	46	-			-	-	-		-		+	+	-
47	47	-	+			-	-				+	+	+
45.8	48	-				-	-			+	+	+	+
45.6	49	-	+			-				+	+	+	+

46	50	-	+			-				+	+	+	+
46.9	51	-	+			-				+	+	+	+
47.8	52	-	+			-				+	+	+	+
48.2	53	-	+			-				+	+	+	+
48.3	54	-	+			-				+	+	+	+
85	55	-				-	-				+	+	+
47.2	56	-	+			-					+	+	+
47.2	57	-				-	-			+	+	+	+
48.1	58	-	+			-				+	+	+	+
49.4	59	-	+			-	-			+	+	+	+
50.6	60	-	+			-	-			+	+	+	+
51.5	61	-	+			-	-			+	+	+	+
51.6	62	-	+			-	-			+	+	+	+
51.2	63	-	+			-	-			+	+	+	+
50.5	64	-	+			-	-			+	+	+	+
50.1	65	-	+			-	-			+	+	+	+
49.8	66	-	+			-	-			+	+	+	+
49.6	67	-	+			-	-			+	+	+	+
49.4	68	-	+			-	-			+	+	+	+
49.3	69	-	+			-	-			+	+	+	+
49.2	70	-	+			-	-			+	+	+	+
49.3	71	-	+			-	-			+	+	+	+
49.7	72	-	+			-	-			+	+	+	+
50.3	73	-	+			-	-			+	+	+	+
51.3	74	-	+			-	-			+	+	+	+
52.8	75	-	+			-	-			+	+	+	+
54.4	76	-	+			-	-			+	+	+	+
56	77	-	+			-	-			+	+	+	+
56.9	78	-	+			-	-			+	+	+	+
57.5	79	-				-	-			+	+	+	+
57.3	80	-				-	-			+	+	+	+
56.6	81	-				-	-			+	+	+	+
56	82	-	+			-	-			+	+	+	+
55.4	83	-	+			-	-			+	+	+	+
55.4	84	-	+			-	-			+	+	+	+
56.4	85	-	+			-	-			+	+	+	+
57.2	86	-	+			-	-			+	+	+	+
58	87	-				-	-			+	+	+	+
58.4	88	-				-	-			+	+	+	+
58.1	89	-				-	-			+	+	+	+
57.5	90	-				-	-			+	+	+	+
56	91	-				-	-			+	+	+	+
54.7	92	-	+			-	-			+	+	+	+
53.2	93	-	+			-	-			+	+	+	+
52.1	94	-	+			-	-			+	+	+	+
51.6	95	-	+			-	-			+	+	+	+
512	96	+	-	-	-	+	+	+	+	-	-	-	-
51.6	97	+	+			+	-			-	+	+	+
50.5	98	-	-	+	+	-	+	+	+	+	-	+	+
51	99	-	+	-		-	-	-		+	+	-	+
51.8	100	-	+		-	-	-		-	+	+	+	-

APPENDIIX II

Table A2: Sign matrix for influence values identifying outliers for Botswana Meat sales data

		ACF Lag (k)				PACF Lag (k)				IACF Lag (k)			
simulated	date	1	2	3	4	1	2	3	4	1	2	3	4
2842	1												
7155	2												
6715	3												
5105	4												
4994	5												
3753	6										+	+	+
6840	7												
5738	8										+	+	+
3437	9											+	+
6508	10												
7756	11												
6522	12												
5532	13												
8604	14		-				-						
8992	15		-				-						
9625	16		-	-	-		-	-	-				
8523	17		-	-	-		-	-	-				
8969	18		-				-	-					
12451	19												
9017	20		-	-	-		-	-	-				
8973	21												
8349	22		-	-	-		-	-	-				
9897	23		-		-		-		-				
5675	24	-				-				-			
7235	25		-	-	-		-	-	-				
24553	26	-				-				-	-	-	-
24106	27	+	+		+	+	+		+	+	-	-	-
15303	28	+	+	+	+	+	+	+	+	+	+	+	+