UNIVERSIDAD COMPLUTENSE
MADRID

COMPLEMENTOS DE FORMACIÓN EN TÉCNICAS DE MINERÍA DE DATOS

Serie Temporal de los números de violaciones en el estado de Rio de Janeiro

Priscilla Toscano Pinel

Octubre/2018

Tabla de Contenido

1. Introducción: Presentación de la serie a analizar.

La serie a ser analizada trata del número de violaciones ocurridas en el estado de Rio de Janeiro (Brasil) de enero de 2006 hasta septiembre de 2018. Los datos fueron obtenidos en la página oficial del Instituto de Segurança Pública (ISP)[1] del estado de Rio de Janeiro, a partir del reporte de "*Séries históricas do estado por mes desde 1991 (números absolutos)".* Es importante destacar que estos números fueron elaborados por la policía civil del estado, a través de denuncias hechas por las víctimas, es decir que la realidad puede ser aún mayor. En estos datos no están solamente violaciones hacia mujeres, pero también a hombres. A finales de febrero y a principios de marzo se celebra el Carnaval en la ciudad.

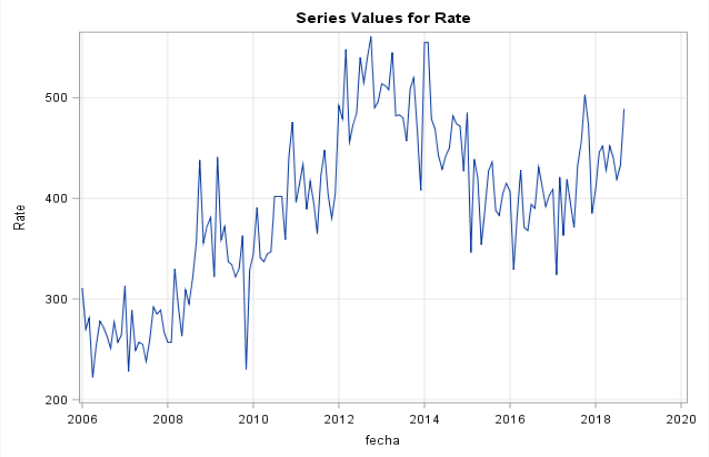
El objetivo del trabajo es entender lo que ocurrirá, en marzo de 2019 con las tasas de violaciones en este estado, si el gobierno no hace ninguna campaña de concientización o empiece algún programa cuya meta sea la reducción de esta tasa.

[1] http://www.ispdados.rj.gov.br/estatistica.html

2. Representación gráfica y descomposición estacional (si tuviera comportamiento estacional).

```
proc timeseries data=series.ISPRJ PLOTS=(DECOMP PERIODOGRAM SERIES)
PRINT=(SEASONS DECOMP);
id FECHA interval=MONTH ;
var RATE;
run;
```

Representación gráfica de la serie



Gráficos de descomposición estacional

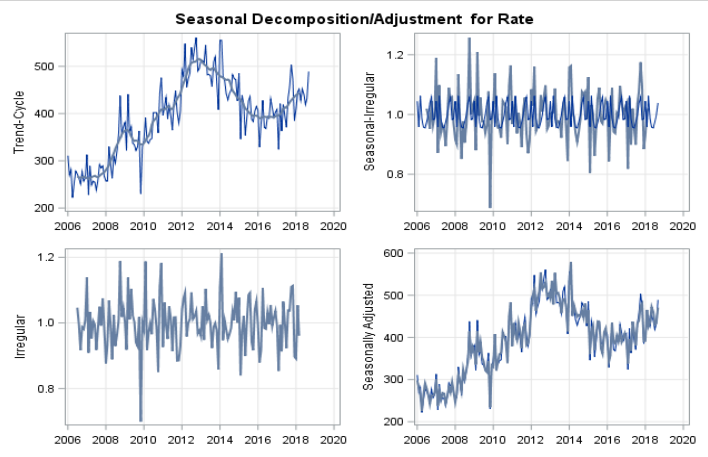


Tabla de descomposición estacional

Seasonal Decomposition for Variable Rate

Obs	Time	Season	Trend-Cycle	Seasonal	Irregular	Seasonally Adjusted
1	JAN2006	1	.	1.044257	.	297.8193
2	FEB2006	2	.	0.958629	.	281.6522
3	MAR2006	3	.	1.062303	.	264.5196
4	APR2006	4	.	0.984966	.	225.3886
5	MAY2006	5	.	0.957539	.	265.2634
6	JUN2006	6	.	0.954904	.	291.1286
7	JUL2006	7	266.7500	0.974276	1.046605	279.1818
8	AUG2006	8	265.0833	0.998326	0.993804	263.4409
9	SEP2006	9	263.6667	1.038594	0.916585	241.6728
10	OCT2006	10	265.1250	1.058505	0.987043	261.6898
11	NOV2006	11	266.3750	0.983018	0.981473	261.4399
12	DEC2006	12	265.5417	0.984682	1.009660	268.1067

Tabla de estadísticos estacionales

Season Statistics for Variable Rate

Season Index	N	Minimum	Maximum	Sum	Mean	Standard Deviation
1	13	257.0000	555.0000	5274.000	405.6923	87.48370
2	13	228.0000	555.0000	4875.000	375.0000	101.60709
3	13	281.0000	548.0000	5343.000	411.0000	82.15940
4	13	222.0000	545.0000	4957.000	381.3077	91.12115
5	13	254.0000	482.0000	4901.000	377.0000	80.41248
6	13	255.0000	485.0000	4909.000	377.6154	72.00410
7	13	238.0000	540.0000	4979.000	383.0000	84.26743
8	13	260.0000	515.0000	5105.000	392.6923	78.03886
9	13	251.0000	540.0000	5375.000	413.4615	87.14606
10	12	277.0000	561.0000	4979.000	414.9167	88.94989
11	12	230.0000	490.0000	4653.000	387.7500	89.01494
12	12	264.0000	496.0000	4646.000	387.1667	71.53745

La serie de violaciones en Rio de Janeiro no es una serie nítidamente estacional, ya que es difícil observar grandes fluctuaciones arriba de la media a un nivel constante a cada año. No obstante, se puede observar un sutil aumento en lo número de violaciones en los meses de marzo, con en mayor coeficiente de estacionalidad de 6,23%, lo que coincide con las fechas de carnaval en este estado. Además, también se puede observar que en los meses de enero (4,42%) y octubre (5,85%) hay más violaciones que en la media, mientras en junio (-4,52%) hay menos violaciones que la media. La máxima fue 561 en octubre de 2012 y la mínima 222 en April 2006. Se observa también una tendencia positiva 2013, y luego una tendencia negativa hasta 2017, donde los números vuelven a subir.

3. Encontrar el modelo de suavizado exponencial más adecuado. Para dicho representar gráficamente la serie observada y la suavizada con las predicciones para un periodo que se considere adecuado.

Para la decisión del método de suavizado más adecuado, se ha hecho un teste para cada método (lineal, double y multwinters), se ha analizado cada uno de sus coeficientes de R^2 y la suma de los cuadrados de los errores (SSE). Abajo se puede ver la tabla de comparación de los resultados obtenidos en cada método. Como el método de multwinters presenta el mayor R^2 y la menos SSE, se queda comprobado que este es el mejor método de suavizado, como ya se era esperado, teniendo en consideración que este método es lo que lleva en consideración la estacionalidad de la serie.

Método Suavizado	Lineal	Doble (Brown)	Multwinters
R^2	0,77631527	0,76790523	0,79539373
SSE	234.404,017	243.217,074	214.411,283

```
proc esm data=series.ISPRJ out=salida lead=12 print=(FORECASTS estimates all)
PLOTS=(FORECASTS MODELS SEASONS);
id FECHA interval=month;
forecast rate / model=MULTwinters;
run;
```

Gráfico la serie observada y la suavizada con las predicciones

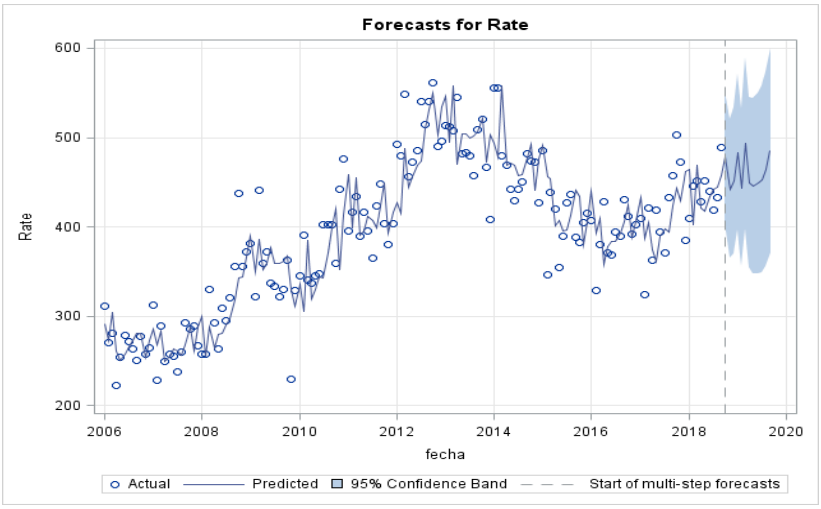


Tabla de Predicciones

Forecasts for Variable Rate					
Obs	Time	Forecasts	Standard Error	95% Confidence Limits	
154	OCT2018	478.9410	37.8075	404.8397	553.0424
155	NOV2018	442.6203	39.6178	364.9709	520.2698
156	DEC2018	451.9655	41.7874	370.0637	533.8673
157	JAN2019	483.3500	44.4772	396.1763	570.5238
158	FEB2019	443.5279	45.0356	355.2598	531.7960
159	MAR2019	493.6149	48.7028	398.1592	589.0707
160	APR2019	449.6952	48.4292	354.7756	544.6148
161	MAY2019	445.6576	49.8730	347.9082	543.4069
162	JUN2019	449.0112	51.6379	347.8028	550.2197
163	JUL2019	452.9739	53.3988	348.3141	557.6337
164	AUG2019	464.3159	55.5991	355.3438	573.2881
165	SEP2019	485.4450	58.4775	370.8312	600.0588

Tabla de estadísticos del modelo

Winters Method (Multiplicative) Parameter Estimates				
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|
Level Weight	0.33939	0.04201	8.08	<.0001
Trend Weight	0.0010000	0.0098306	0.10	0.9191
Seasonal Weight	0.09756	0.04618	2.11	0.0363

La estimación del parámetro de suavizado (α) vale 0.339, que es mayor que 0, y por lo tanto podemos aceptarlo para el modelo. Lo mismo pasa con el coeficiente de estacionalidad, que rechazamos la hipótesis que sea nulo. Sin embargo, el coeficiente de tendencia(β) asume un valor muy próximo a cero, y con eso aceptamos la hipótesis nula para el coeficiente de tendencia.

4. Representar la serie y los correlogramas. Decidir qué modelo puede ser ajustado. Ajustar el modelo adecuado comprobando que sus residuales están incorrelados. (Sintaxis, tablas de los parámetros estimados y gráficos)

```
proc timeseries data=series.ISPRJ PLOTS=(SERIES ACF PACF)OUTCORR=AUTOCOR PRINT=ALL;
id FECHA interval=MONTH ;
var RATE;
run;
```

Desde del grafico de correlación simples, se puede observar que la serie decrece de forma lenta, la media no es constante, y por lo tanto **no se trata de una serie estacionaria**. El coeficiente de autocorrelación simples de orden 1 está relativamente cerca de 1 (0.8366290373), es decir, que esa variable explica bien la tasa, y son dependientes. Esto nos permite interpretar que el tiempo es una buena variable para explicar la tasa de violaciones, es decir, que se en un mes ha porque ha tenido muchas violaciones, es probable que en próximo mes va a tener más violaciones este mes.

Gráfico de correlación simples

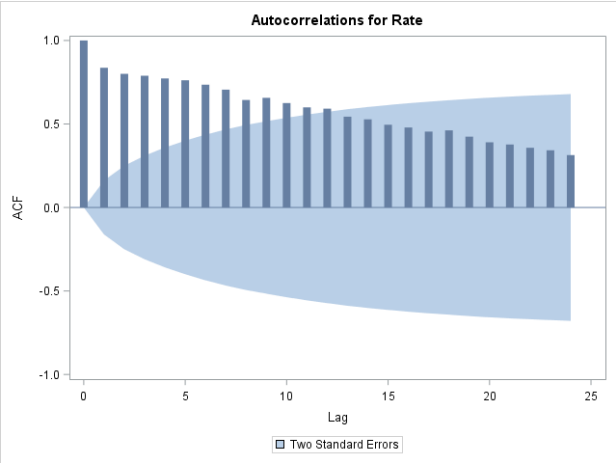
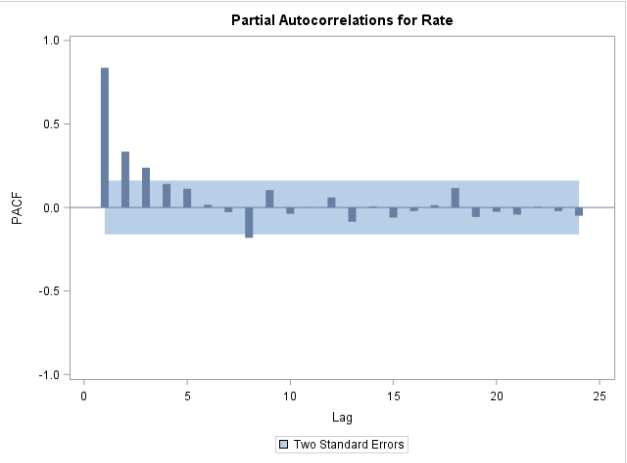


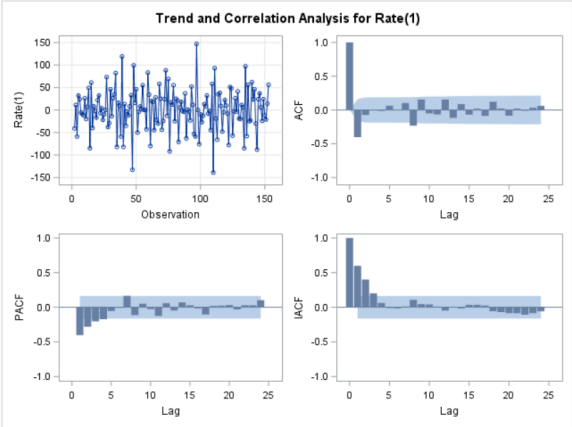
Gráfico de correlación parcial



	Variable Name	Time Lag	Number of Variance Products	Autocovariance	Autocorrelations	Autocorrelation Standard Errors	Partial Autocorrelations	Partial Autocorrelation Standard Errors	Inverse Autocorrelations	Inverse Autocorrelation Standard Errors	White Noise Test Statistics	White Noise Test Probabilities
1	Rate	0	153	6849.1593831	1	0	.	.	1	.	.	.
2	Rate	1	152	5730.2056212	0.8366290373	0.080845208	0.8366290373	0.080845208	-0.271403731	0.080845208	109.2057256	0
3	Rate	2	151	5480.6526452	0.8001934746	0.125242152	0.3340933484	0.080845208	-0.020935904	0.080845208	209.7682559	0
4	Rate	3	150	5403.0905702	0.7888691543	0.155098861	0.2384287383	0.080845208	-0.079184418	0.080845208	308.1561851	0
5	Rate	4	149	5292.241127	0.7726847677	0.179417077	0.1412366716	0.080845208	-0.062156791	0.080845208	403.1819912	0
6	Rate	5	148	5217.4554201	0.7617658063	0.199987386	0.1117117812	0.080845208	-0.073900169	0.080845208	496.1651643	5.37013E-105
7	Rate	6	147	5035.8135039	0.7352454837	0.218129337	0.0162300865	0.080845208	-0.028227522	0.080845208	583.3760155	8.98034E-123
8	Rate	7	146	4832.1608227	0.7055115164	0.233766743	-0.028263828	0.080845208	-0.073982899	0.080845208	664.2257295	3.54916E-139
9	Rate	8	145	4408.9528423	0.6437217468	0.247292095	-0.181731542	0.080845208	0.2014498187	0.080845208	731.9979147	9.21564E-153
10	Rate	9	144	4497.1801218	0.6566032224	0.258011787	0.1038423648	0.080845208	-0.062414007	0.080845208	802.9992729	4.80806E-167
11	Rate	10	143	4282.4483257	0.6252516676	0.268711257	-0.03804347	0.080845208	-0.040026653	0.080845208	867.832376	5.32131E-180
12	Rate	11	142	4106.2190714	0.5995216116	0.278057659	-0.001032399	0.080845208	0.0463836399	0.080845208	927.8590763	6.34399E-192
13	Rate	12	141	4053.2044354	0.5917812988	0.286381639	0.059770328	0.080845208	-0.118943089	0.080845208	986.7605946	1.31462E-203
14	Rate	13	140	3723.9636043	0.5437110448	0.294265675	-0.085380881	0.080845208	0.0867211089	0.080845208	1036.836799	2.13745E-213
15	Rate	14	139	3614.7342218	0.5277631925	0.300760076	0.0059745123	0.080845208	-0.042150391	0.080845208	1084.357903	1.22186E-222
16	Rate	15	138	3393.0856843	0.4954017704	0.306753306	-0.06046699	0.080845208	0.0398839892	0.080845208	1126.533298	9.74861E-231
17	Rate	16	137	3280.7820996	0.4790050744	0.311938669	-0.021847587	0.080845208	0.0387213341	0.080845208	1166.250879	2.60003E-238
18	Rate	17	136	3112.4740295	0.4544315376	0.316709687	0.0133447831	0.080845208	0.05972719	0.080845208	1202.260715	4.29454E-245
19	Rate	18	135	3162.5020266	0.4617357912	0.320943105	0.1165223403	0.080845208	-0.125088731	0.080845208	1239.712838	3.45317E-252
20	Rate	19	134	2904.8591707	0.4241190792	0.325255906	-0.056605867	0.080845208	0.0179728995	0.080845208	1271.547039	4.40923E-258
21	Rate	20	133	2672.7208643	0.3902261161	0.328850631	-0.025993688	0.080845208	-0.013775581	0.080845208	1298.699184	5.6189E-263
22	Rate	21	132	2582.4276175	0.3770430024	0.331863344	-0.042706766	0.080845208	0.0244349436	0.080845208	1324.239777	1.56969E-267
23	Rate	22	131	2448.2605188	0.3574541607	0.334651458	0.003751819	0.080845208	-0.019830112	0.080845208	1347.370678	1.42386E-271
24	Rate	23	130	2345.5841714	0.3424630732	0.337137716	-0.021890111	0.080845208	-0.016223693	0.080845208	1368.76543	3.00734E-275
25	Rate	24	129	2148.2525206	0.3136519973	0.339403777	-0.04934583	0.080845208	0.0353204967	0.080845208	1386.850889	3.20577E-278

Después de muchos intentos, lo mejor modelo que fue encontrado fue un Modelo Arima (0,1,1), una vez que la serie no posé una fuerte estacionalidad, una diferenciac

```
Modelo ARIMA (0,1,1)
PROC ARIMA DATA= SERIES.ISPRJ;
IDENTIFY VAR=rate (1) nlag=10;
ESTIMATE q=1 noconstant;
run;
```

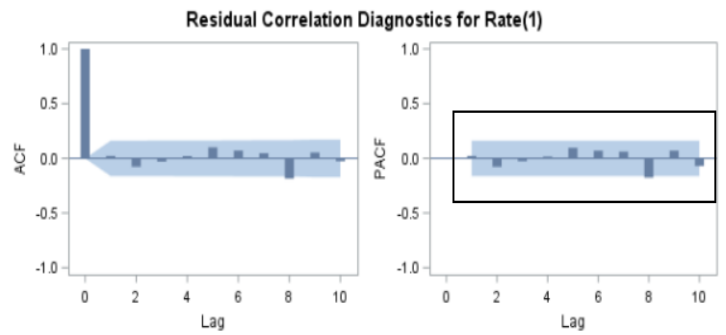


Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.64280	0.06255	10.28	<.0001	1

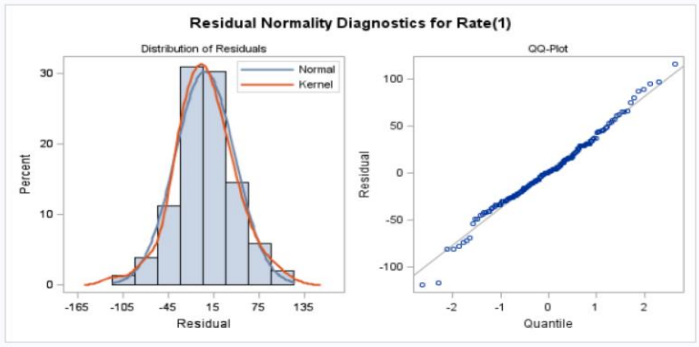
El para el modelo se ha utilizado **noconstant**, una vez que el valor de sin ello, tenía un Pvalor de MU(0,3446) mayor que 0,3446> α. Así que seguimos con la análisis sin la constante.

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	3.79	5	0.5794	0.026	-0.072	-0.024	0.026	0.105	0.076
12	12.27	11	0.3435	0.052	-0.179	0.058	-0.020	-0.017	0.113
18	16.99	17	0.4550	-0.054	0.023	-0.065	-0.026	-0.042	0.131
24	21.62	23	0.5433	0.053	-0.038	0.021	0.043	0.098	0.097
30	29.95	29	0.4166	-0.021	0.114	-0.073	-0.143	0.017	-0.070

Primeramente, para la decisión si el modelo es bueno, se analizan los residuos con el contraste de **Ljung-box sobre las autocorrelaciones**, donde en este caso, todos los contrastes de aleatoriedad de los residuales tienen los p-valores mayores que α (nivel de significación igual a 0,05) y s que no están en torno al 0.05, aceptando que estos son aleatorios, por lo tanto, los residuos están incorrelados, siendo el ajuste del modelo bueno.



Ahora, seguimos con la analice de los coeficientes de autocorrelación de los residuos, y se observa que ellos están dentro de las bandas de confianza, es decir, no podemos rechazar que ninguno de ellos individualmente sea distinto de cero.



Al analizar el gráfico de la normalidad de los residuos, se observa que estos tienen un comportamiento normal.

5. Escribir la expresión algebraica del modelo ajustado con los parámetros estimados.

Moving Average Factors	
Factor 1:	1 - 0.6428 B**(1)

$$(1-\Phi_1 B^s-\Phi_2 B^{2s}-\ldots-\Phi_P B^{Ps})(1-\phi_1 B-\phi_2 B^2-\ldots-\phi_p B^p)(1-B^s)^D(1-B)^d X_t=(1-\Theta_1 B-\Theta_2 B^{2s}-\ldots-\Theta_Q B^{Qs})(1-\theta_1 B-\theta_2 B^2-\ldots-\theta_q B^q)Z_t$$

P=0

$$(1-B).X_t=(1-\theta B).Z_t$$

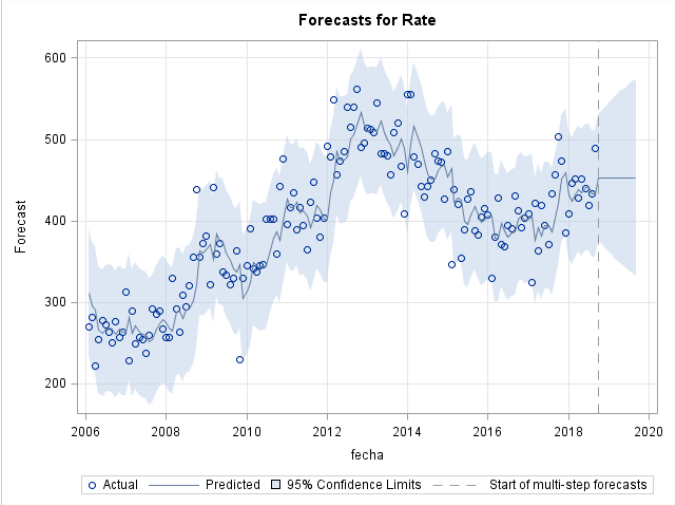
$$X_t-X_{t-1}=Z_{t-1}-0{,}6428B.Z_t$$

$$X_t=X_{t-1}+Z_{t-1}+0{,}6428.Z_t$$

6. Calcular las predicciones y los intervalos de confianza para las unidades de tiempo que se considere oportuno, dependiendo de la serie, siguientes al último valor observado. Representarlas gráficamente.

```
PROC ARIMA data=series.isprj2 plots=all;
identify var=rate (1) ;
estimate Q=1 noconstant;
forecast lead=12 id=fecha interval=month out=PREDICCIONES2 Printall;
run;
```

Después de numerosos intentos por ajustar las predicciones, con la serie entera, con menos datos, sin la constante, seguimos observando que el modelo ARIMA(0,1,1) asume predicciones en forma lineal. Por ejemplo, para el mes de marzo de 2019, que es el Carnaval en la ciudad, la predicción para la tasa de violación sería de 452,99, con un intervalo de confianza de 353,5 y 552,4.



The SAS System

Obs	fecha	Rate	FORECAST	STD	L95	U95	RESIDUAL
154	01OCT18	.	452.994	39.6591	375.264	530.725	.
155	01NOV18	.	452.994	42.1133	370.453	535.535	.
156	01DEC18	.	452.994	44.4322	365.909	540.079	.
157	01JAN19	.	452.994	46.6358	361.589	544.399	.
158	01FEB19	.	452.994	48.7400	357.465	548.523	.
159	01MAR19	.	452.994	50.7570	353.512	552.476	.
160	01APR19	.	452.994	52.6969	349.710	556.278	.
161	01MAY19	.	452.994	54.5678	346.043	559.945	.
162	01JUN19	.	452.994	56.3767	342.498	563.490	.
163	01JUL19	.	452.994	58.1293	339.063	566.925	.
164	01AUG19	.	452.994	59.8306	335.728	570.260	.
165	01SEP19	.	452.994	61.4848	332.486	573.502	.

7. Conclusiones

Comparando los modelos de predicciones, sacando como ejemplo la fecha de marzo de 2018, donde la media era de 411 violaciones (o sea que a cada día más de 13 personas son violadas en Rio), con el modelo de winters, se predice una 493 (con un STD de 48,7) y el modelo ARIMA (0,1,1) 455 (STD=50,7). Este estudo nos permite observar que para este caso, el modelo suavizado de winters no da unas predicción más segura que el modelo ARIMA. Sin embargo, es importante resaltar que este modelo ARIMA (0,1,1) fue el mejor encontrado en numerosas tentativas, pero que quizás había un todavía mejor, que nos daría una mejor predicción.

Por fin, con cualquier modelo de predicción se puede confirmar que la tasa de violación en RJ va aumentar, si el gobierno local no reacciona. Con eso, queda claro la importancia y urgencia de que el gobierno empiece un programa que evite que casi aproximadamente 493 personas sean violadas en el próximo carnaval y en cualquier fecha.