

DESEMPREGO REGISTRADO NO ALGARVE

PREVISÃO DE UMA SÉRIE SAZONAL COM BASE EM MODELOS UNIVARIADOS

Ricardo Martinho

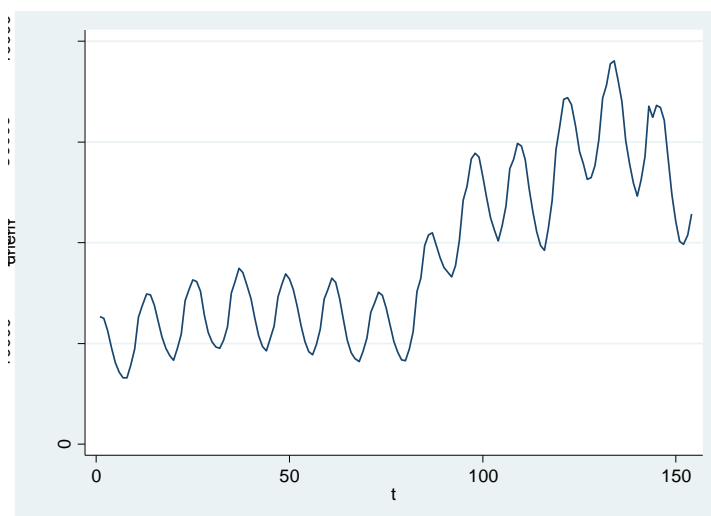
09/01/2015

Definição do problema, descrição dos dados e breve revisão da literatura

Neste trabalho pretende-se modelar o comportamento do desemprego no Algarve, uma região cuja actividade económica tem um forte carácter associado ao turismo, com o objectivo de obter uma previsão para a evolução futura do nível de desemprego, num horizonte temporal de 12 meses, com base em modelos univariados, determinísticos e estocásticos. Este pode ser um tema de interesse para entidades, como o Instituto Emprego e Formação Profissional (IEFP), para avaliar a necessidades de formação disponíveis para os desempregados ao longo do ano, ou como a Segurança Social para prever os meses em que terá maior despesa relacionada com o pagamento dos subsídios de desemprego. Os dados utilizados estão disponíveis desde Janeiro de 2002 no *site* do IEFP e dizem apenas respeito aos indivíduos inscritos no Centro de Emprego, deixando eventualmente de fora desempregados da economia paralela, ou desencorajados que deixaram de procurar emprego activamente. Em particular, neste trabalho utilizaram-se 154 observações mensais, correspondentes ao período entre Janeiro de 2002 e Outubro de 2014, sendo que para estimação dos modelos apenas foram tidos em conta dados até Dezembro de 2013. Existe uma extensa literatura sobre modelos para previsão do nível de desemprego. Destaque para os modelos multivariados que exploram a relação entre o desemprego e outras variáveis macroeconómicas como o PIB, o investimento empresarial, incluindo factores institucionais como alterações de fiscalidade ou de salário mínimo previstas. Outras abordagens tiram partido da informação contida em indicadores avançados de alta frequência como a confiança na indústria, a confiança dos consumidores ou as encomendas de matérias-primas. Em todo o caso, existe também alguma prevalência de modelos matemáticos-estatísticos mais puros, alguns complementando as dinâmicas provenientes dos modelos ARMA/ARIMA com componentes ARCH/GARCH e outros utilizando modelos de espaço de estados ou modelos de redes neuronais.

Caracterização da série a modelar

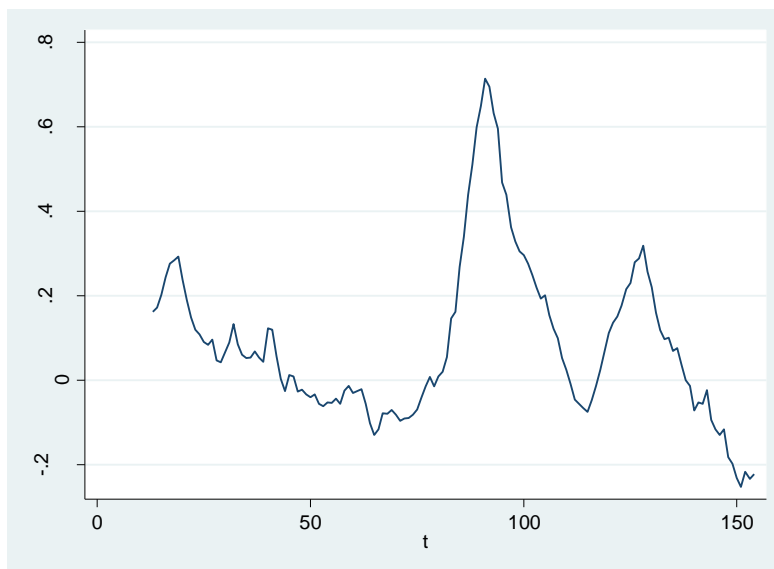
Gráfico 1. Y_t : Desemprego registado no Algarve (Janeiro de 2002 - Outubro de 2014).



A série apresenta uma evolução sazonal muito marcada, com os picos de desemprego a registarem-se nos meses de Dezembro/Janeiro, os mínimos nos meses de Julho/Agosto e um perfil monotónico entre os extremos relativos da série. Adicionalmente, a análise gráfica sugere uma série não estacionária em média e, ainda que de forma menos clara, não estacionária em variância. O correlograma da série original (*ver Anexo 1, Fig 1*) revela uma elevada persistência, com um coeficiente de autocorrelação no primeiro *lag* próximo de 1 e com um decaimento bastante lento e oscilatório, consistente com o carácter sazonal. O resultado do teste de Dickey-Fuller para raiz unitária não permite rejeitar a hipótese nula (de existência de uma raiz unitária) para os níveis de significância convencionais, quer na versão simples, quer nas versões com tendência ou incluindo desfasamentos da própria variável (*Fig 2, anexo*).

Neste contexto, numa primeira fase, procedeu-se à estabilização da variância através da transformação logarítmica (de base natural) e, subsequentemente, à aplicação de uma diferença sazonal com objectivo de resolver o problema da elevada persistência. A série obtida, que pode ser interpretada como taxa de variação homóloga do desemprego, continua a ser não-estacionária em média (gráfico abaixo) e a exibir uma elevada persistência (Anexo, Fig. 3) e uma raiz unitária. O correlograma continua a evidenciar elevada persistência, com a autocorrelação no primeiro *lag* próxima de 1 e um decaimento lento da FAC, agora sem o carácter oscilatório. O teste de Dickey-Fuller continua a não rejeitar a hipótese nula de existência de raiz unitária (Anexo, Fig.4).

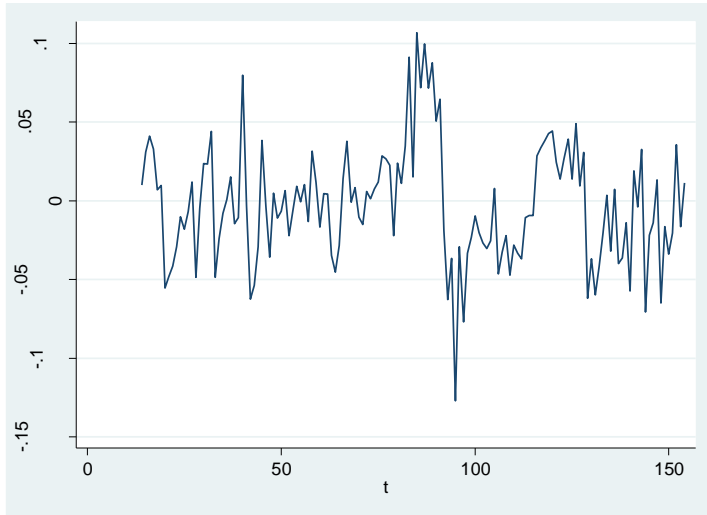
Gráfico 2. $Z_t = (1 - B^{12}) \ln(Y_t)$



Após a aplicação de uma diferença simples adicional obteve-se a série transformada: $X_t = (1 - B)(1 - B^{12}) \ln(Y_t)$, que pode ser interpretada como a variação mensal na taxa de variação homóloga do desemprego. Neste caso, os testes Dickey-Fuller rejeitam a hipótese de raiz unitária (*Anexo, Fig. 6*). O correlograma da série transformada X_t é apresentado na *Fig. 5* em anexo. A FAC apresenta um decaimento amortecido para zero, sugerindo que um modelo univariado do tipo (S)Arima incorpore uma componente autoregressiva. Por sua vez a FACP apresenta um queda para zero a partir do 3º *lag* (inclusive) o que é consistente com um AR(2).

Adicionalmente, observa-se um aumento da FAC e da FACP no primeiro desfaseamento sazonal, sugerindo a incorporação de um termo autoregressivo ou de médias móveis na componente sazonal, não sendo contudo óbvio, pela evolução da FAC e da FACP, qual o mais adequado.

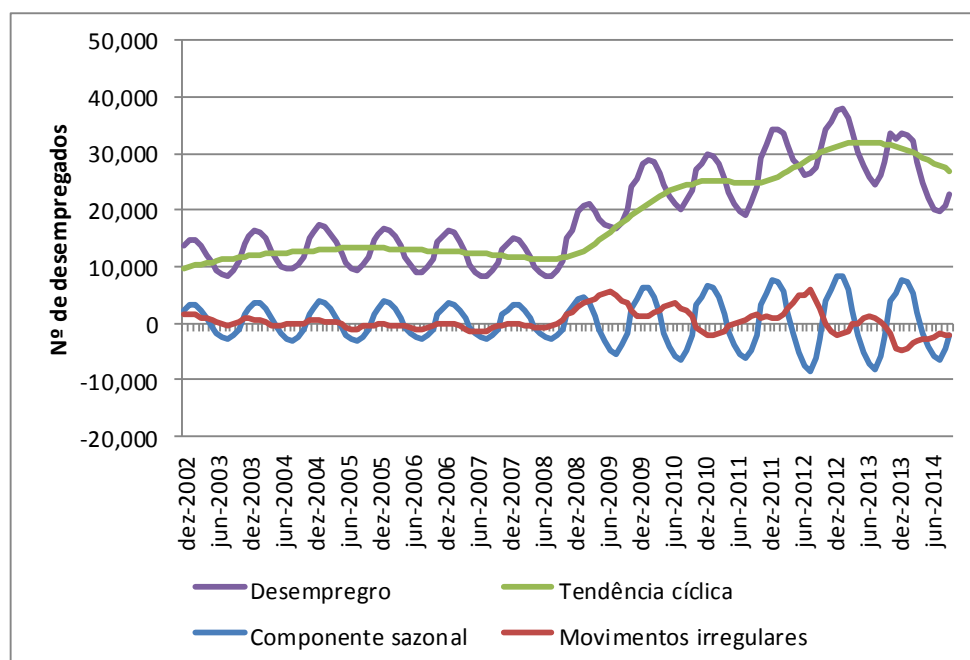
Gráfico 3. $X_t=(1-B)(1-B^{12})\ln(Y_t)$



Decomposição das forças componentes

A decomposição da série em tendência cíclica, componente sazonal e movimentos irregulares foi efectuada pelos métodos multiplicativo dado que a componente sazonal da série aparenta variar com o nível da série. A tendência cíclica foi obtida através de uma média móvel não centrada a 12 meses. No período em análise, destaque-se a existência de alterações significativas na tendência da série, bem como a intensificação do carácter sazonal e a maior magnitude dos movimentos irregulares.

Gráfico 4. Modelo multiplicativo: $Y_t=f(TC_t S_t E_t)= TC_t \times S_t \times E_t$



Métodos de alisamento exponencial (H-W multiplicativo)

Dada a tendência apresentada e o carácter sazonal que varia com o nível da série, foi estimado um modelo Holt-Winters multiplicativo, definido pelas seguintes equações de actualização:

$$a(t) = \alpha Y_t / S(t-s) + (1-\alpha) [a(t-1) + b(t-1)], \quad 0 < \alpha < 1$$

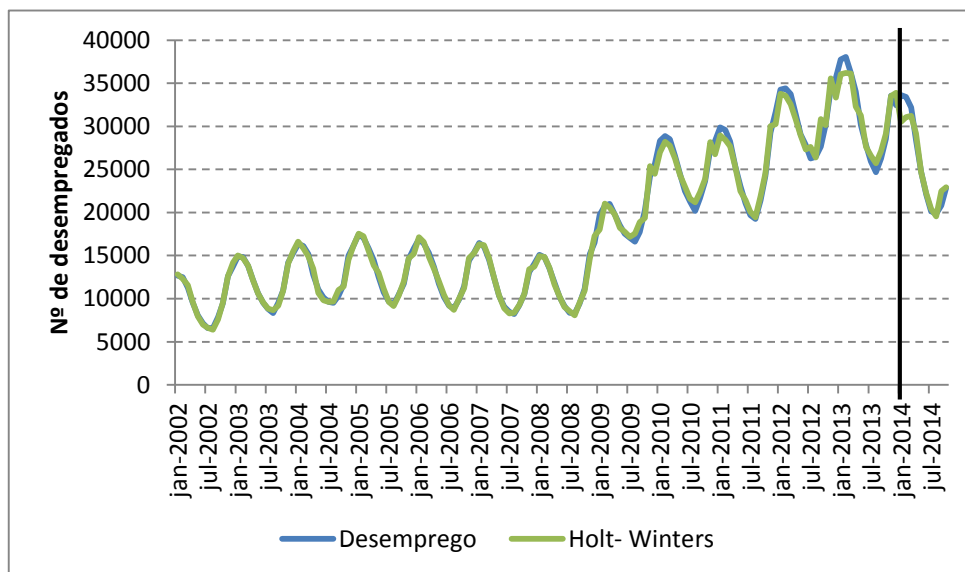
$$b(t) = \beta [a(t) - a(t-1)] + (1-\beta) b(t-1), \quad 0 < \beta < 1$$

$$S(t) = \gamma Y(t) / a(t) + (1-\gamma) S(t-s), \quad 0 < \gamma < 1$$

$$P_{t+h} = [a(t) + b(t) \times h] \times S(t+h-s)$$

A estimação que teve em consideração 144 observações até dez-2013, resultando em $\alpha=0.9289$, $\beta=1$, $\gamma=0.77$ e com $RMSE=678.3094$. O gráfico 5 mostra as previsões *in-sample* obtidas com base neste método até dez-2013, sendo que a partir de Jan-2014 os valores dizem respeito a previsões *out-of-sample*, um passo à frente.

Gráfico 5. Série obtida pelo método de alisamento exponencial H-W sazonal.



Modelos estocásticos(S)ARMA/(S)ARIMA

De acordo com a caracterização da desenvolvida anteriormente foi estimado inicialmente um modelo SARIMA (2,1,0) (0,1,1)¹² com os seguintes resultados:

Quadro 1. Resultados da estimação do modelo proposto pela análise do correlograma da série transformada X_t . SARIMA (2,1,0) (0,1,1)¹²

ARIMA regression

Sample: 14 - 144	Number of obs	=	131
	Wald chi2(3)	=	117.03
Log likelihood = 286.3673	Prob > chi2	=	0.0000

DS12.		OPG				
log_unem	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
log_unem _cons	-.0012558	.0039067	-0.32	0.748	-.0089127	.0064012
ARMA						
ar						
L1.	.4603017	.0911651	5.05	0.000	.2816214	.638982
L2.	.2404494	.0701824	3.43	0.001	.1028945	.3780043
ARMA12						
ma						
L1.	-.601133	.0927707	-6.48	0.000	-.7829602	-.4193058
/sigma	.0265838	.0014108	18.84	0.000	.0238187	.0293489

Note: The test of the variance against zero is one sided, and the two-sided confidence interval is truncated at zero.

Os parâmetros associados aos termos autoregressivos são estatisticamente significativos para um nível de significância de 1% em ambos os casos, assim como o parâmetro da componente de médias móveis sazonal. O correlograma dos resíduos (*Quadro 2*) confirma que estamos na presença de um ruído branco, com o *p-value* de 0.4435 (44%) para o teste de significação global da FAC dos resíduos, e, portanto, na presença de um modelo com as propriedades estatísticas desejáveis.

Quadro 2. Correlograma dos resíduos do modelo SARIMA (2,1,0) (0,1,1)¹²

. corrgram erro_log_unem_arima_is_1

LAG	AC	PAC	Q	Prob>Q	-1	0	1	-1	0	1
					[Autocorrelation]			[Partial Autocor]		
1	0.0448	0.0479	.26897	0.6040						
2	0.0984	0.1032	1.5757	0.4548						
3	-0.0410	-0.0530	1.8041	0.6140						
4	-0.0357	-0.0438	1.9793	0.7396						
5	-0.0088	0.0039	1.99	0.8505						
6	-0.1993	-0.2092	7.5271	0.2748	—			—		
7	-0.2053	-0.2201	13.451	0.0619	—			—		
8	0.0262	0.0843	13.548	0.0943						
9	0.0699	0.1167	14.245	0.1139						
10	-0.0640	-0.1519	14.835	0.1382					—	
11	0.0841	0.0981	15.86	0.1464						
12	-0.0016	0.0023	15.861	0.1977						
13	0.1726	0.0704	20.26	0.0889	—					
14	0.1243	0.1251	22.561	0.0678					—	
15	0.0081	0.0595	22.571	0.0937						
16	0.0055	-0.0238	22.576	0.1256						
17	-0.0733	-0.1007	23.397	0.1368						
18	-0.0933	-0.0854	24.74	0.1323						
19	-0.1402	-0.1091	27.799	0.0874	—					
20	-0.1365	-0.1136	30.725	0.0589	—					
21	0.0155	0.1417	30.763	0.0777					—	
22	0.0723	0.0756	31.599	0.0845						
23	-0.0100	-0.0953	31.615	0.1084						
24	-0.0562	-0.1671	32.129	0.1238					—	
25	0.0273	-0.0414	32.252	0.1509						
26	0.0185	-0.0418	32.309	0.1831						
27	0.0034	-0.0006	32.311	0.2208						
28	-0.0588	0.0275	32.895	0.2396						
29	-0.0016	0.0181	32.896	0.2820						
30	-0.0001	-0.1293	32.896	0.3271					—	
31	-0.0567	-0.0871	33.456	0.3489						
32	-0.0188	0.0964	33.518	0.3936						
33	-0.0327	0.0141	33.708	0.4331						
34	0.0204	0.0334	33.783	0.4782						
35	0.0632	0.1263	34.508	0.4917					—	
36	0.0257	-0.0612	34.629	0.5338						
37	0.1451	0.1508	38.532	0.4001	—				—	
38	0.0993	0.1439	40.38	0.3656					—	
39	-0.0341	-0.1520	40.6	0.3997					—	
40	-0.0057	-0.0730	40.606	0.4435					—	

.

Adicionalmente, foram estimados diversos modelos na vizinhança do proposto com base na análise do correlograma, isto é, com +1/-1 termo autoregressivo ou de médias móveis. O quadro resumo da qualidade de ajustamento dos diversos modelos é apresentado abaixo.

Quadro 3. Comparação da qualidade de ajustamento in-sample

	AIC	BIC	loglikelihood
SARIMA (2,1,0) (0,1,1) ¹²	-562.735	-548.359	286.3673
SARIMA (3,1,0) (0,1,1) ¹²	-563.794	-546.543	287.897
SARIMA (2,1,0) (0,1,0) ¹²	-531.849	-520.348	269.9246
SARIMA (1,1,0) (0,1,0) ¹²	-523.48	-514.854	264.7398
SARIMA (0,1,1) (0,1,0) ¹²	-505.233	-496.607	255.6163
SARIMA (0,1,0) (0,1,0) ¹²	-482.603	-476.853	243.3016

O modelo SARIMA (2,1,0) (0,1,1)¹² sugerido pelo autocorrelograma é o que minimiza o critério de informação de Schwartz (BIC). Por sua vez, o modelo que inclui mais um termo autoregressivo, SARIMA (3,1,0) (0,1,1)¹² é o que apresenta o menor critério de Akaike. De um conjunto de outros modelos considerados (não apresentados no quadro), concluiu-se que alguns modelos com mais parâmetros apresentavam um valor superior da função de verosimilhança, mas o acréscimo residual não justificava a sua adopção com base nos critérios de informação.

A detecção de *outliers* foi efectuada recorrendo ao *software* TRAMO, disponível no Eviews. Foi detectado um único *outlier* do tipo aditivo (AO) em Novembro de 2008. O modelo escolhido pelo TRAMO foi o SARIMA (1,1,0) (0,1,1)¹² com $\phi=-0.6408$ e $\Theta=-0.609$, ambos significativos para um nível de significância de 1%.

Os modelos de intervenção/detecção de *outliers* são úteis para melhorar a precisão das estimativas após uma intervenção. No entanto, têm pouca utilidade na previsão *out-of-sample*, na medida em que não é possível estimar os parâmetros do modelo de intervenção antes da mesma ocorrer.

Previsão out-of-sample

Muitas vezes, mais do que escolher um modelo com boa qualidade de ajustamento e propriedades desejáveis do ponto de vista estatístico (e.g. ausência de autocorrelação nos resíduos) é importante escolher um modelo que funcione bem fora da amostra. Uma necessidade de muitas entidades é a de obter, no início do ano, uma projecção para a evolução de uma variável até ao final do ano, i.e. previsões $P_t(1)$ até $P_t(12)$. Neste contexto, foram estimadas versões sucessivas dos 6 modelos (S)ARIMA identificados anteriormente, a cada 12 meses, para incorporar as “novas” observações disponíveis. A primeira ronda foi estimada com observações até dez-2007¹, a partir da qual se efectuaram previsões *out-of-sample* de a 1 até 12 passos à frente. As previsões obtidas foram confrontadas com os valores reais ao longo de 2008, tendo sido calculados os respectivos erros de previsão. O processo foi

¹ Para que existisse um mínimo de 60 observações na primeira estimação. No limite este processo poderia ser repetido mês após mês, mesmo com o objectivo de prever 1 a 12 passos à frente. Os modelos de ajustamento exponencial não foram incluídos neste exercício pelo facto de o STATA não conseguir convergir para uma solução em algumas iterações do exercício.

depois repetido para observações até dez-2009, dez-2010,... dez-2013². O modelo com melhor capacidade de previsão no período em análise é o SARIMA (2,1,0) (0,1,0)¹², enquanto que o modelo com melhor ajustamento *in-sample*, identificado anteriormente, é o SARIMA (2,1,0) (0,1,1)¹², de acordo com BIC, e o SARIMA (3,1,0) (0,1,1)¹², de acordo com o AIC.

Tal como no estudo da série não sazonal, existe heterogeneidade observada ao longo do tempo, com diferentes modelos a apresentarem a melhor performance em diferentes anos. Focando agora a análise no período completo, concluiu-se, através da ferramenta de optimização do Excel, que existe uma combinação de dois dos seis modelos apresentados, com pesos inversamente proporcionais aos respectivos erros quadráticos médios, que supera os resultados dos modelos individuais. Trata-se da previsão combinada dos modelos SARIMA (2,1,0) (0,1,0)¹² e SARIMA (2,1,0) (0,1,1)¹². Refira-se ainda que, um exercício análogo, avaliado à luz do Erro Absoluto Médio, produziu exactamente o mesmo resultado para o conjunto do período (2008-2014).

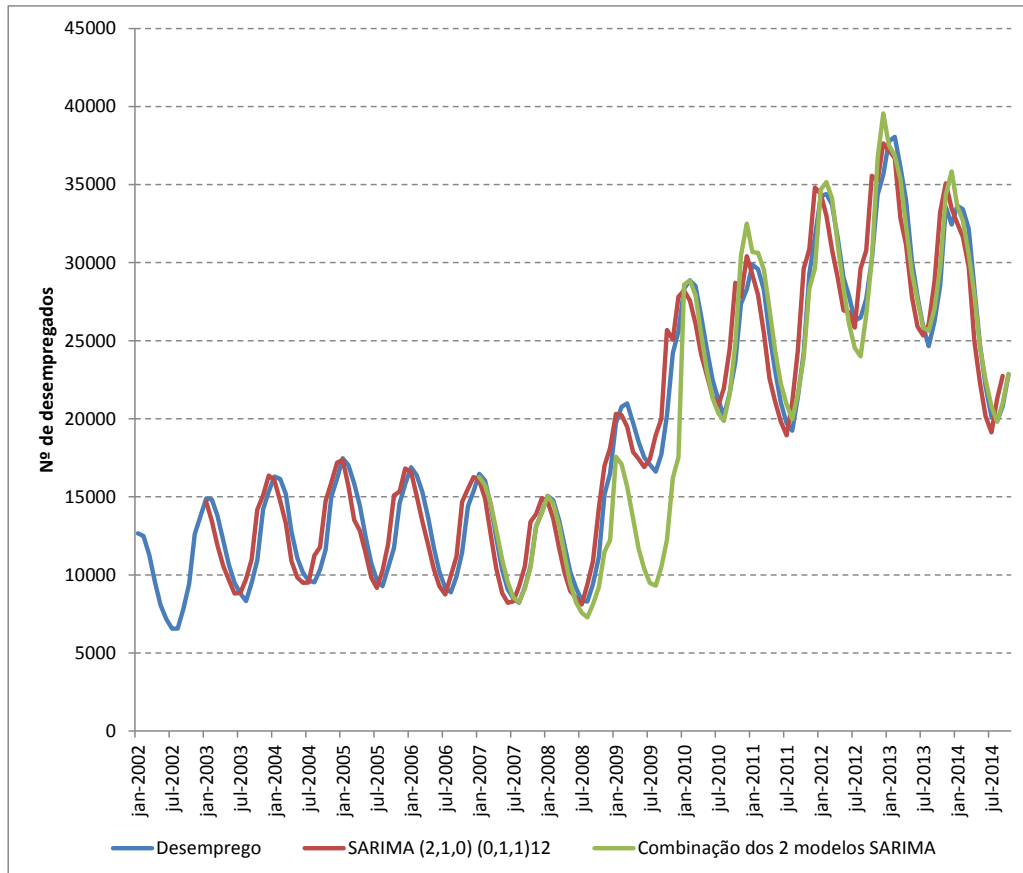
Quadro 4. Comparação das previsões *out-of-sample*. De 1 a 12 passos à frente. Erro quadrático médio.

RMSE	SARIMA(2,1,0)(0,1,1)12 dynamic	SARIMA(2,1,0)(0,1,0)12 dynamic	SARIMA(1,1,0)(0,1,0)12 dynamic	SARIMA(0,1,1)(0,1,0)12 dynamic	SARIMA(0,1,0)(0,1,0)12 dynamic	SARIMA(3,1,0)(0,1,1)12 dynamic	Combinação 1+2
2008	1826	1856	1842	1834	1885	1836	1840
2009	7150	6276	5932	6168	6046	7094	6694
2010	3653	4337	6061	7399	7139	3550	1695
2011	741	1919	2260	2621	2776	547	1163
2012	1655	1842	1972	2840	3151	1661	1732
2013	1541	1228	2513	3794	4027	2148	1318
2014	916	692	922	1584	2220	1031	675
(2008-2014)	3077	3004	3416	4039	4064	3092	2711

Por fim apresenta-se um gráfico comparativo entre a série original, a série com as previsões *in-sample* do modelo escolhido pelo critério de informação BIC e a série com as previsões *out-of-sample* resultante da combinação óptima dos dois modelos. De facto, observa-se que o erro de projecção é mais elevado no caso do exercício de estimação de 1 até 12 passos à frente, e que o erro tende a aumentar de mês para mês, dentro do ano de projecção. Em todo o caso, a existência de uma estrutura sazonal marcada, permite que se obtenham previsões *out-of-sample* razoáveis para o nível do desemprego no Algarve com base nos valores da própria série (modelos univariados).

² Para 2014 apenas foram calculados em erros de previsão até Setembro.

Gráfico 6. Série original, previsões in-sample e combinação previsões out-of-sample, 1 a 12 passos à frente.



Anexo 1

Fig. 1

. corrgram unem

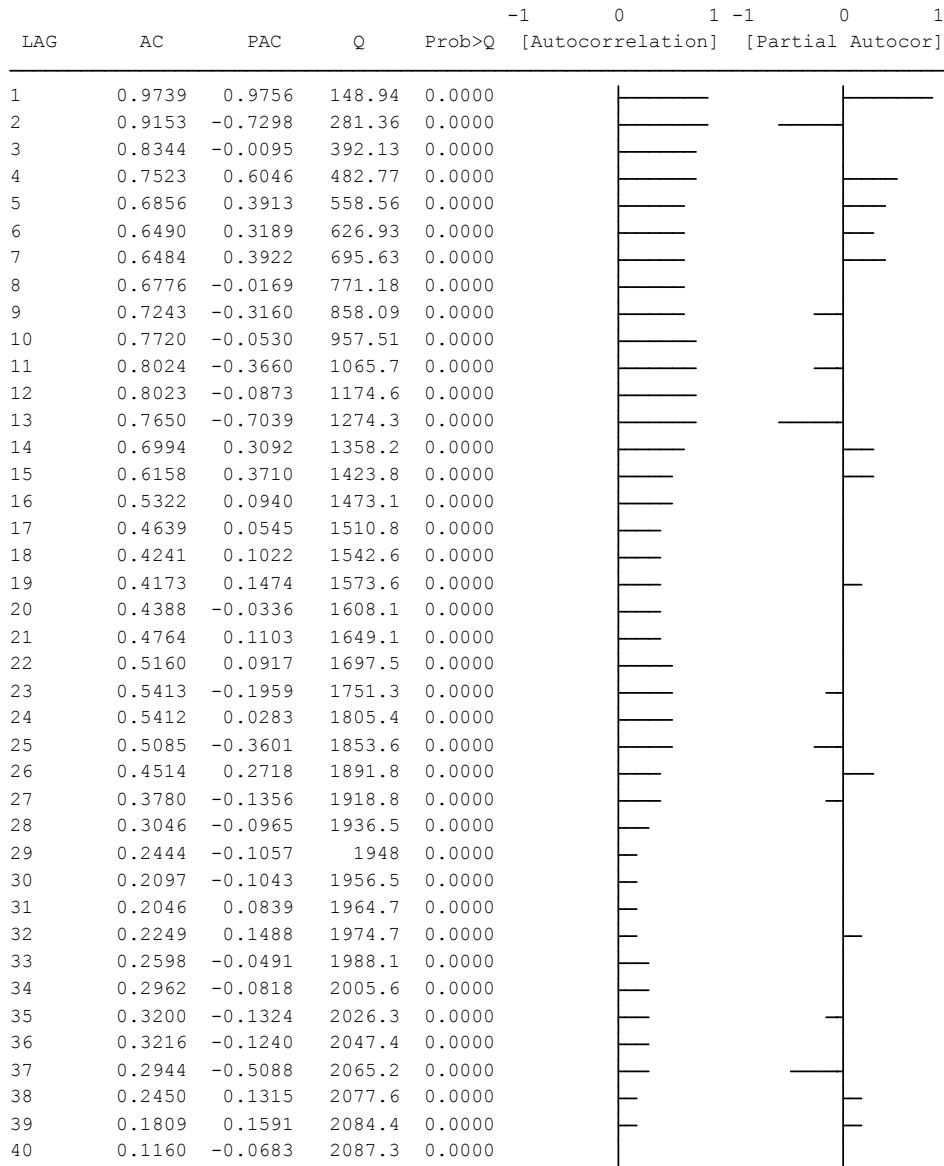


Fig. 2

```
. dfuller unem, lags(0)

Dickey-Fuller test for unit root                Number of obs   =       153

          Test          _____ Interpolated Dickey-Fuller _____
          Statistic      1% Critical   5% Critical   10% Critical
                        Value          Value          Value
-----
Z(t)          -1.388          -3.492          -2.886          -2.576
-----
MacKinnon approximate p-value for Z(t) = 0.5882

. dfuller unem, lags(12)

Augmented Dickey-Fuller test for unit root      Number of obs   =       141

          Test          _____ Interpolated Dickey-Fuller _____
          Statistic      1% Critical   5% Critical   10% Critical
                        Value          Value          Value
-----
Z(t)          -2.733          -3.496          -2.887          -2.577
-----
MacKinnon approximate p-value for Z(t) = 0.0685

. dfuller unem, trend lags(0)

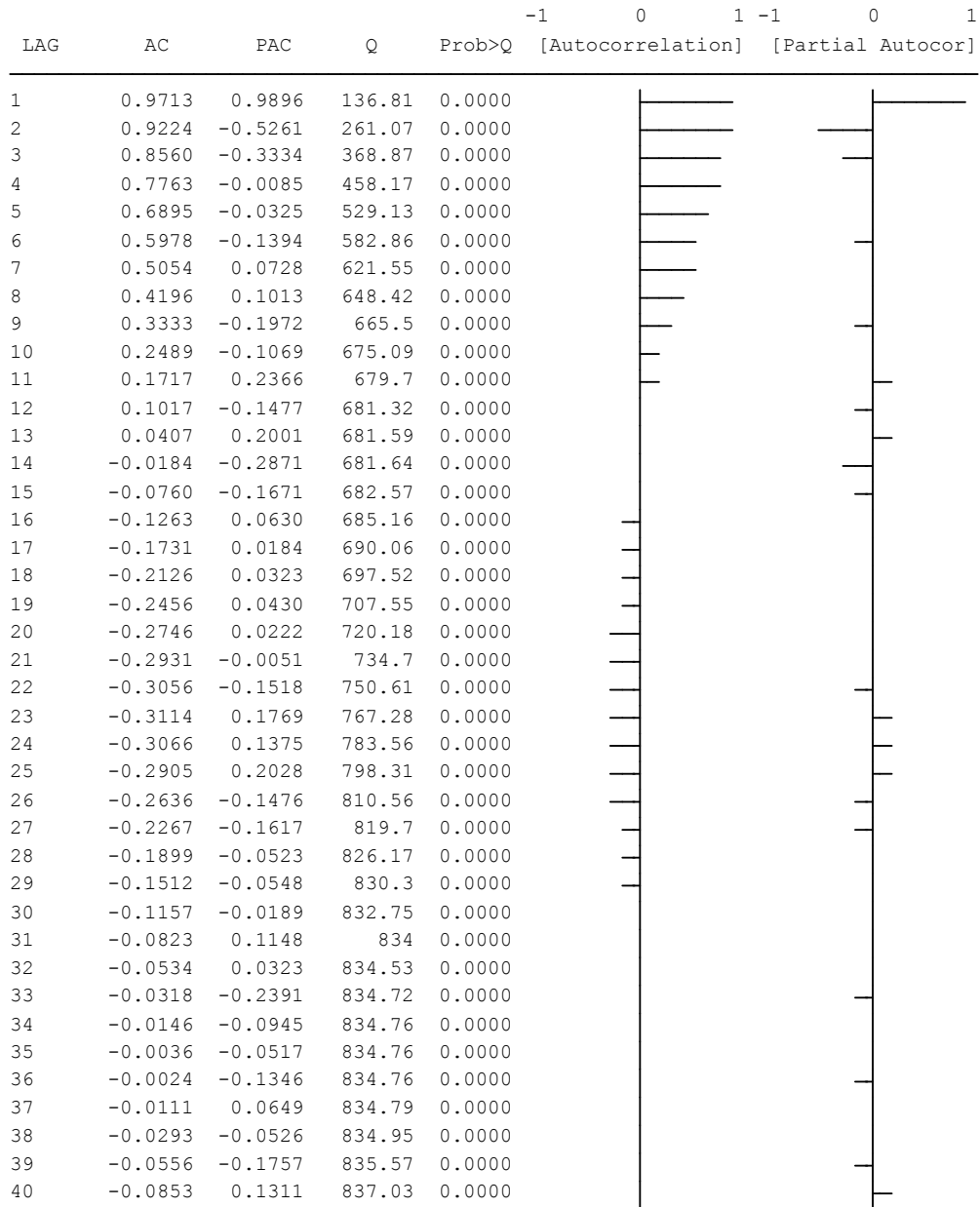
Dickey-Fuller test for unit root                Number of obs   =       153

          Test          _____ Interpolated Dickey-Fuller _____
          Statistic      1% Critical   5% Critical   10% Critical
                        Value          Value          Value
-----
Z(t)          -2.241          -4.022          -3.443          -3.143
-----
MacKinnon approximate p-value for Z(t) = 0.4667

.
.
end of do-file
```

Fig. 3

. corrgram S12.log_unem



.

Fig 4

```
. dfuller S12.log_unem, lags(0)

Dickey-Fuller test for unit root          Number of obs   =      141

          Test          _____ Interpolated Dickey-Fuller _____
          Statistic      1% Critical   5% Critical   10% Critical
                          Value         Value         Value
-----
Z(t)          -0.623          -3.496          -2.887          -2.577
-----
MacKinnon approximate p-value for Z(t) = 0.8657

. dfuller S12.log_unem, lags(12)

Augmented Dickey-Fuller test for unit root      Number of obs   =      129

          Test          _____ Interpolated Dickey-Fuller _____
          Statistic      1% Critical   5% Critical   10% Critical
                          Value         Value         Value
-----
Z(t)          -1.722          -3.500          -2.888          -2.578
-----
MacKinnon approximate p-value for Z(t) = 0.4200

. dfuller S12.log_unem, trend lags(0)

Dickey-Fuller test for unit root          Number of obs   =      141

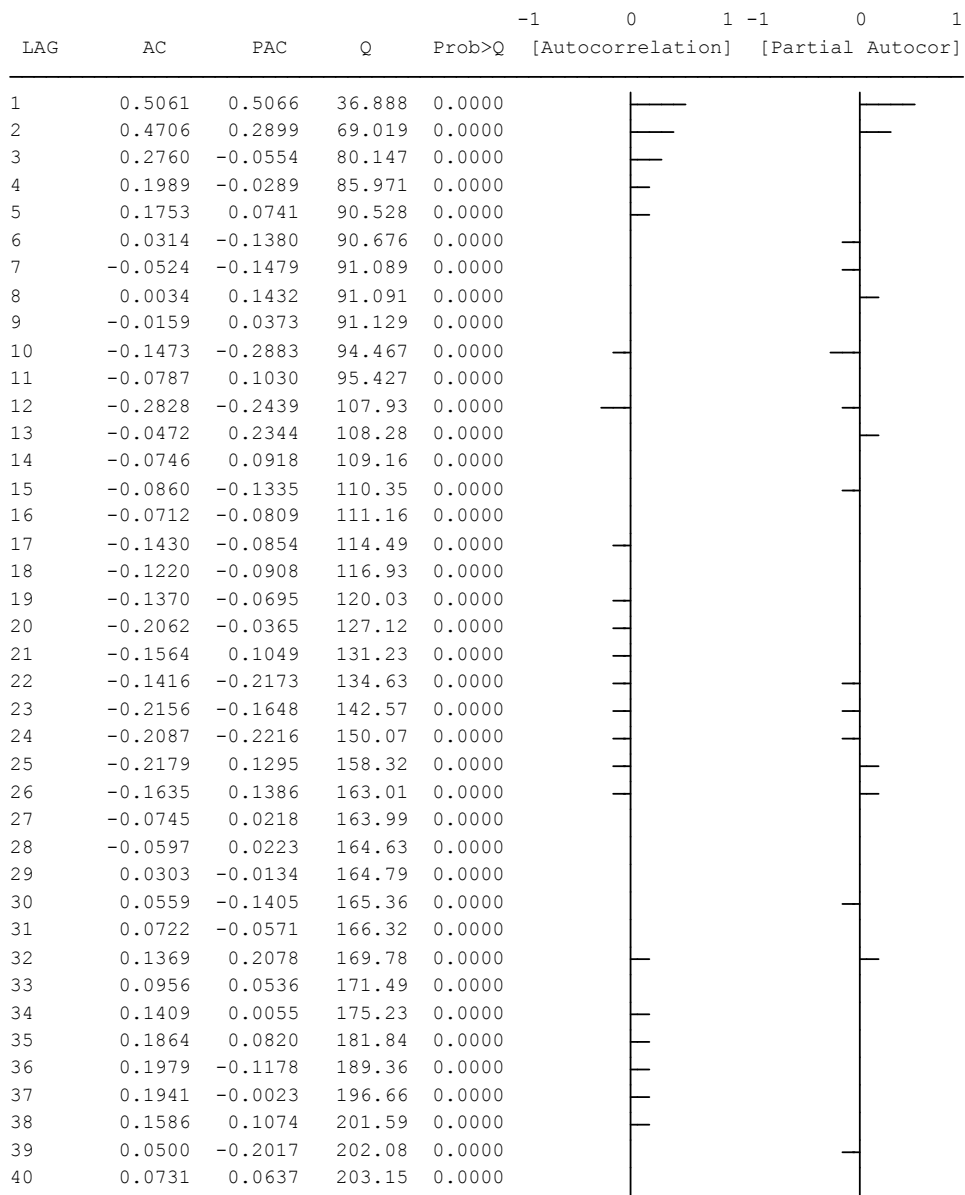
          Test          _____ Interpolated Dickey-Fuller _____
          Statistic      1% Critical   5% Critical   10% Critical
                          Value         Value         Value
-----
Z(t)          -0.701          -4.026          -3.445          -3.145
-----
MacKinnon approximate p-value for Z(t) = 0.9731

.
.
end of do-file

.
```

Fig 5

. corrgram DS12.log_unem



.

Fig 6

```
. dfuller DS12.log_unem, lags(0)

Dickey-Fuller test for unit root          Number of obs   =       140

          Test          _____ Interpolated Dickey-Fuller _____
          Statistic      1% Critical   5% Critical   10% Critical
                          Value         Value         Value
-----
Z(t)          -6.722          -3.497          -2.887          -2.577
-----
MacKinnon approximate p-value for Z(t) = 0.0000

. dfuller DS12.log_unem, lags(12)

Augmented Dickey-Fuller test for unit root      Number of obs   =       128

          Test          _____ Interpolated Dickey-Fuller _____
          Statistic      1% Critical   5% Critical   10% Critical
                          Value         Value         Value
-----
Z(t)          -2.928          -3.501          -2.888          -2.578
-----
MacKinnon approximate p-value for Z(t) = 0.0422

. dfuller DS12.log_unem, trend lags(0)

Dickey-Fuller test for unit root          Number of obs   =       140

          Test          _____ Interpolated Dickey-Fuller _____
          Statistic      1% Critical   5% Critical   10% Critical
                          Value         Value         Value
-----
Z(t)          -6.715          -4.027          -3.445          -3.145
-----
MacKinnon approximate p-value for Z(t) = 0.0000

.
.
end of do-file

.
```