4
Estudo de casos

4.1.
Introdução

Para avaliar os conceitos explorados nos capítulos anteriores, dois estudos de caso foram selecionados: (i) **derivados do petróleo** e (ii) **competição NN3**. No primeiro, são analisadas séries de vendas de dois produtos importantes no mercado brasileiro de derivados – óleo diesel e gás liquefeito de petróleo. No segundo, utiliza-se um banco de dados com 11 séries relacionadas a atividades de transporte, originalmente distribuídas aos participantes da competição NN3 – uma competição organizada pela comunidade científica para avaliar a eficácia de metodologias de previsão baseadas em inteligência computacional.

Todos os modelos, algoritmos e testes de hipótese utilizados neste trabalho foram implementados em plataforma Windows/PC, com os seguintes programas: MATLAB (Mathworks, 2010c), Forecast Pro (BFS, 2010) e Excel (Microsoft, 2011).

4.2.
Previsores disponíveis (*experts*)

Qualquer método de combinação pressupõe a existência de previsores individuais. Neste trabalho, os previsores disponíveis são derivados das metodologias listadas a seguir, detalhadas no Apêndice A. O processo de escolha dos previsores seguiu dois critérios: (i) terem naturezas diferentes entre si, buscando **complementaridade**, e (ii) serem capazes de representar tendência e sazonalidade, características recorrentes na maioria das séries tratadas.

1. Holt-Winters multiplicativo (HW);
2. Regressão harmônica (REG);
3. Decomposição clássica (DEC);
4. ARIMA Box & Jenkins (BJ).

4.3.
Metodologia de avaliação dos resultados

4.3.1.
Métricas de desempenho

A principal métrica de desempenho utilizada neste trabalho foi o SMAPE – *Symmetric Mean Absolute Percentage Error* (73). O SMAPE, medido em pontos percentuais (pp), é uma métrica amplamente utilizada em trabalhos relacionados a séries temporais, principalmente por seu emprego em competições para modelos de previsão (Makridakis & Hibon, 2000; NN3, 2011).

$$SMAPE = \frac{1}{H}\sum_{h=1}^{H}\frac{\left|y_{\tau+h} - \hat{y}_{\tau+h|\tau}\right|}{\left(\frac{\left|y_{\tau+h}\right| + \left|\hat{y}_{\tau+h|\tau}\right|}{2}\right)}100\% \tag{73}$$

Outras métricas eventualmente utilizadas foram o RAE – *Relative Absolute Error* e o coeficiente UTHEIL. As equações (74) a (76) definem estas métricas.

$$RAE = \frac{1}{H}\sum_{h=1}^{H}\frac{\left|y_{\tau+h} - \hat{y}_{\tau+h|\tau}\right|}{\left|y_{\tau+h} - \mu_y\right|} \tag{74}$$

onde

$$\mu_y = E(y_t) \tag{75}$$

$$t < \tau$$

$$UTHEIL = \sqrt{\frac{\sum_{h=1}^{H}(y_{\tau+h} - \hat{y}_{\tau+h|\tau})^2}{\sum_{h=1}^{H}(y_{\tau+h} - y_{\tau})^2}} \tag{76}$$

As métricas RAE (Witten & Frank, 2005) e UTHEIL (Makridakis et al., 1998) fornecem importantes medidas de desempenho relativo. A primeira compara a previsão do modelo testado com uma previsão simplória, igual à média da série dentro da amostra; a segunda, estabelece o mesmo tipo de comparação, mas ao invés da média da amostra, utiliza como referência um valor fixo igual à

última realização da série no conjunto de treinamento – o que se chama de previsão **ingênua**. Tanto o RAE quanto o UTHEIL devem, idealmente, apresentar valores menores do que 1.

4.3.2. Testes de hipótese

Para garantir a validade das conclusões tomadas, empregou-se uma bateria de testes de hipótese sobre os desempenhos dos métodos de combinação utilizados neste trabalho, sempre com nível de significância de 5%. Os testes escolhidos – **teste t, teste de sinais** e **teste de Wilcoxon** (Kachigan, 1986; Flores, 1986, 1989; Gibbons, 1992) – verificam se a mediana[11] das diferenças de desempenho entre dois métodos é (estatisticamente) nula. Como será visto mais adiante, propõe-se uma arquitetura de comparação onde as diferenças de desempenho são medidas para cada um dos horizontes de previsão considerados (de 1 até *H* passos a frente). A unidade de medida das diferenças é, logicamente, a mesma dos desempenhos sendo comparados (e.g. SMAPE).

O teste t (Kachigan, 1986) é paramétrico: assume que a distribuição das diferenças de desempenho é normal; isto nem sempre pode ser assumido, principalmente se o tamanho da amostra é reduzido (<30). A validade da premissa de distribuição normal pode ser checada pela análise de gráficos Q-Q (Johnson & Wichern, 2007) ou por testes específicos de normalidade, por exemplo, o Jarque-Bera (Cromwell et al. 1994).

Os testes de sinais e de Wilcoxon (*ranking* sinalizado) são não paramétricos, dispensando assunção de normalidade (Gibbons, 1992). Estes testes foram utilizados para comparar métodos de previsão por Flores (1986, 1989).

4.4. CASO 1: Derivados do petróleo

Para as empresas integradas de Petróleo & Gás, o uso de técnicas de séries temporais é útil nas atividades de planejamento relacionadas ao marketing e comercialização de derivados (*downstream*). Não obstante, pode-se encontrar

[11] A **mediana** é considerada uma medida mais robusta (resistente a *outliers*) do que a **média**.

aplicações destas técnicas em outras áreas: financeira, materiais, Gás & Energia e Exploração & Produção (*upstream*).

Com o objetivo de estudar a aplicação das combinações de previsores ao mercado nacional de derivados do petróleo, este trabalho utiliza dados reais publicados pela Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP), relativos às vendas (em MMm³) de **óleo diesel** (DIESEL) e **gás liquefeito de petróleo** (GLP) na região sudeste do país (ANP, 2011.). As Figuras 15 e 16 exibem estas séries.

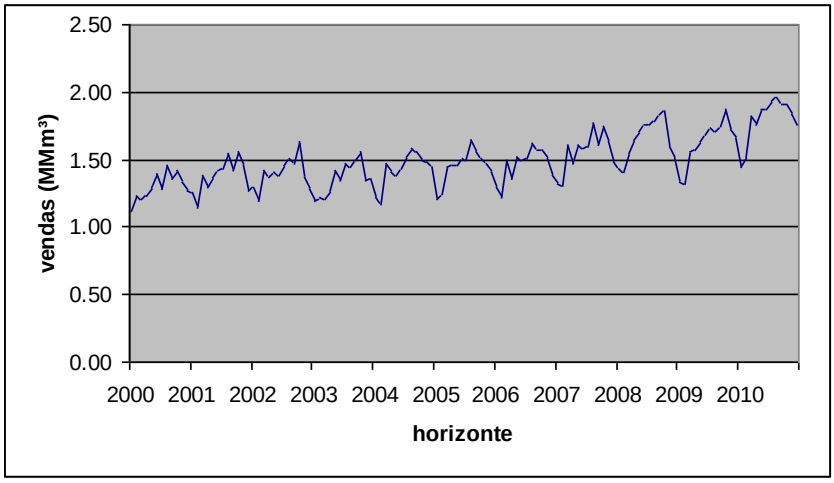

Figura 15 – Vendas de óleo diesel de jan/2000 a dez/2010 na região sudeste. Fonte: ANP.

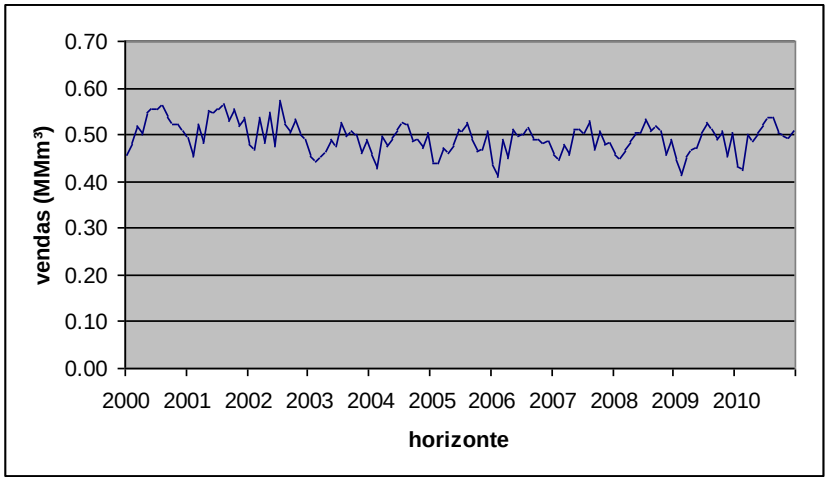

Figura 16 – Vendas de GLP de jan/2000 a dez/2010 na região sudeste. Fonte: ANP.

4.4.1.
Previsores individuais

As metodologias de previsão citadas na seção 4.2 foram aplicadas às séries disponíveis (DIESEL e GLP), separando-se sempre os últimos **12** meses de dados para teste; os modelos resultantes podem ser vistos no Apêndice B. Com os modelos ajustados, foram geradas previsões até 12 passos a frente, de maneira **não recursiva**, i.e., sem reestimação de parâmetros a cada passo. A Tabela 9 exibe os desempenhos **totais**, obtidos dentro e fora da amostra (12 meses).

Tabela 9 – Desempenhos totais

	DIESEL		*GLP*	
Método	*SMAPE Amostra*	*SMAPE Teste*	*SMAPE Amostra*	*SMAPE Teste*
HW	3.52	5.56	3.12	4.18
REG	3.40	7.97	2.98	5.50
DEC	3.27	7.86	2.96	5.57
BJ	3.75	5.18	3.04	3.39

Tanto para DIESEL quanto para GLP, o previsor BJ apresentou o menor **erro composto** (SMAPE amostra + SMAPE teste). As Figuras 17 e 18 exibem as previsões individuais fora da amostra, ao longo do horizonte de 12 meses.

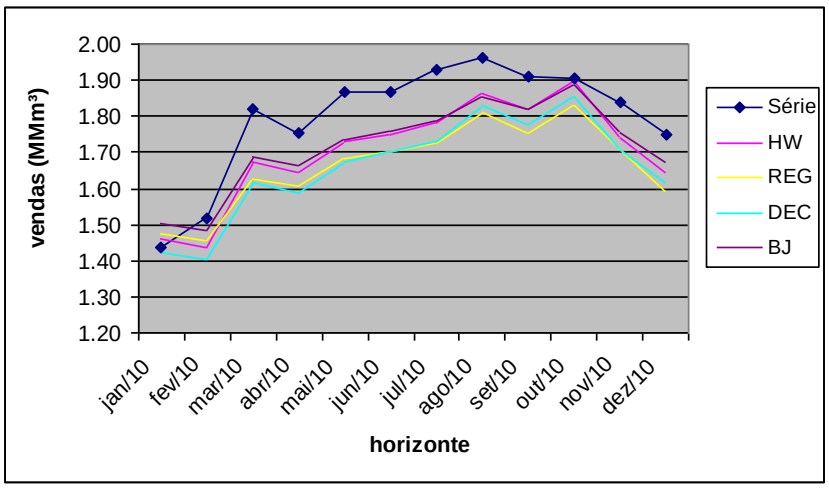

Figura 17 – Previsões para DIESEL (geradas em dez/2009).

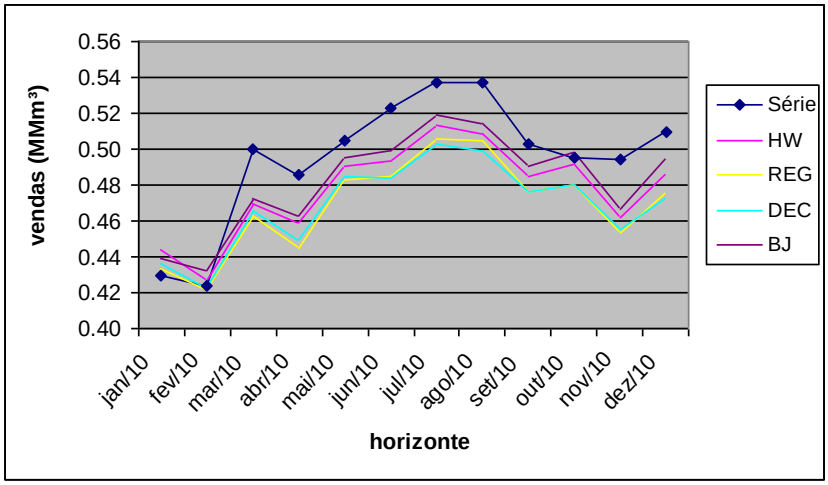

Figura 18 – Previsões para GLP (geradas em dez/2009).

As Tabelas 10 e 11 e as Figuras 19 e 20 exibem a evolução dos SMAPEs ao longo do horizonte de teste (*h*). Como cada SMAPE é uma **média acumulada**, as últimas linhas das Tabelas 10 e 11 equivalem exatamente aos desempenhos **totais**, exibidos na Tabela 9[12]. De maneira geral, os desempenhos para GLP foram melhores do que os desempenhos para DIESEL.

Tabela 10 – Evolução dos SMAPEs fora da amostra (DIESEL)

h	*HW*	*REG*	*DEC*	*BJ*
1	1.61	2.56	1.01	4.46
2	3.47	3.36	4.34	3.29
3	5.11	6.01	6.85	4.80
4	5.46	6.68	7.62	4.96
5	5.92	7.40	8.27	5.43
6	6.05	7.74	8.47	5.54
7	6.32	8.22	8.81	5.85
8	6.18	8.20	8.59	5.82
9	6.02	8.24	8.42	5.70
10	5.47	7.82	7.86	5.23
11	5.48	7.80	7.83	5.20
12	5.56	7.97	7.86	5.18

[12] Esta observação vale para diversas tabelas ao longo deste capítulo.

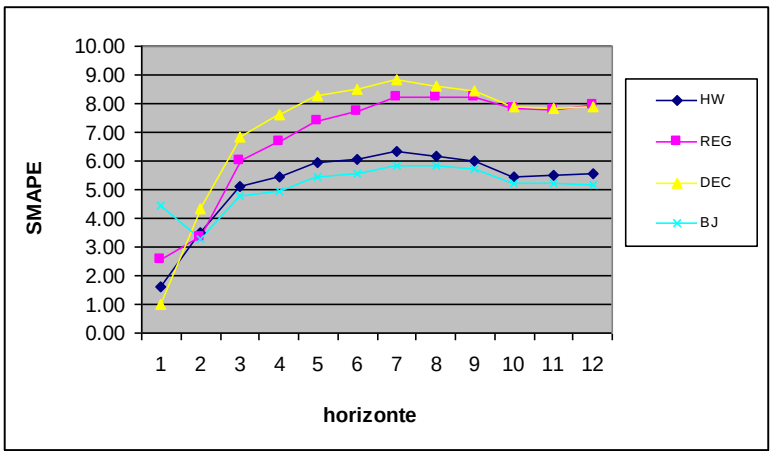

Figura 19 – Evolução dos SMAPEs fora da amostra (DIESEL).

Tabela 11 – Evolução dos SMAPEs fora da amostra (GLP)

h	*HW*	*REG*	*DEC*	*BJ*
1	3.17	0.75	1.40	2.10
2	1.87	0.63	0.91	2.04
3	3.34	2.98	3.00	3.26
4	3.92	4.41	4.18	3.64
5	3.72	4.42	4.13	3.30
6	4.05	4.93	4.73	3.52
7	4.13	5.08	5.00	3.51
8	4.31	5.23	5.31	3.61
9	4.24	5.26	5.34	3.49
10	3.88	5.05	5.11	3.21
11	4.14	5.37	5.40	3.44
12	4.18	5.50	5.57	3.39

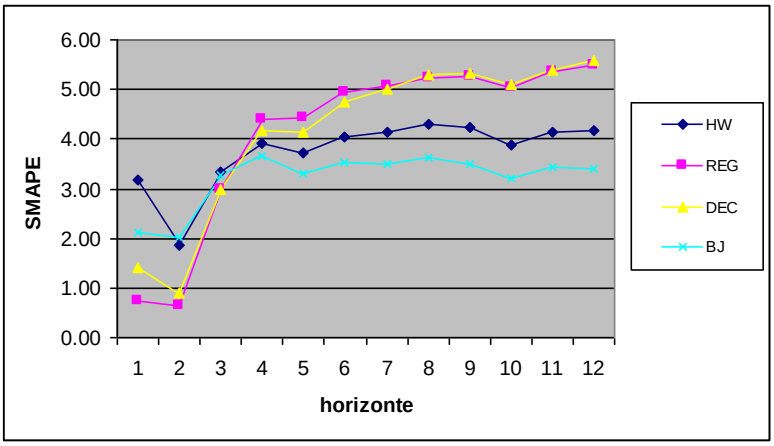

Figura 20 – Evolução dos SMAPEs fora da amostra (GLP).

4.4.2.
Combinações tradicionais

4.4.2.1.
Experimentos

Antes de passar à metodologia NEW, é fundamental que se avalie o funcionamento das combinações tradicionais, principalmente para que se estabeleça um patamar comparativo. Nesta linha, foi definida a sequência de experimentos da Tabela 12, organizada de acordo com os previsores sendo combinados; como se observa, todas as possíveis combinações foram testadas.

As 11 combinações da Tabela 12 foram avaliadas com os seguintes métodos para geração de pesos (seção 2.4) – média simples (AVG), mínimo quadrados irrestritos (MQI), mínimos quadrados restritos (MQR), Bates & Granger simples (BG1), Bates & Granger correlacionado (BG2), AFTER (AFTER), *shrinkage* MQI (SMQI), *shrinkage* MQR (SMQR), *shrinkage* BG1 (SBG1), *shrinkage* BG2 (SBG2) e *shrinkage* AFTER (SAFTER). Como pode ser observado na seção 2.4, excetuando-se a média simples, a utilização padrão (*default*) dos métodos supracitados gera pesos de maneira **dinâmica**: os vetores calculados podem variar ao longo do horizonte de previsão.

Tabela 12 – Experimentos de combinação tradicional

Experimento	*Previsores Combinados*
2a	HW, REG
2b	HW, DEC
2c	HW, BJ
2d	REG, DEC
2e	REG, BJ
2f	DEC, BJ
3a	HW, REG, DEC
3b	HW, REG, BJ
3c	HW, DEC, BJ
3d	REG, DEC, BJ
4a	HW, REG, DEC, BJ

Considerando apenas o número de combinações (onze) e a quantidade de métodos de geração de pesos (onze), ter-se-ia um total de 121 experimentos por série. Contudo, dois outros parâmetros foram checados.

Primeiro, utilizou-se duas janelas de tempo nos métodos de geração: janela **mínima**, i.e., de tamanho (*v*) igual ao número de previsores na combinação, e janela **expansiva** (seção 2.4). O termo **janela mínima** foi utilizado pelo fato dos métodos MQI e MQR só serem computáveis para janelas maiores ou iguais ao número de previsores sendo combinados (*N*). Depois, para cada tamanho de janela, testou-se a possibilidade de combinação **estática**: manutenção do mesmo vetor de pesos gerado no último instante do histórico por todo o horizonte de previsão. Assim, o total de experimentos subiu para **484** por série.

As Tabelas 13 a 16 exibem os resultados selecionados, obtidos nos conjuntos de **validação** (constituído pelos últimos 12 meses dentro da amostra, imediatamente anteriores ao conjunto de teste) e **teste** (12 meses fora da amostra). Em todos os casos, o método de geração de pesos foi escolhido pelo melhor desempenho na validação; os valores em negrito indicam os experimentos com menor erro composto (SMAPE validação + SMAPE teste).

Tabela 13 – Desempenhos totais para *janela mínima & pesos dinâmicos*

	DIESEL			*GLP*		
Experimento	*Método*	*SMAPE Validação*	*SMAPE Teste*	*Método*	*SMAPE Validação*	*SMAPE Teste*
2a	BG1	3.24	6.89	SBG1	2.55	4.79
2b	MQR	2.50	7.71	SBG1	2.62	4.80
2c	MQR	4.31	5.20	AFTER	2.54	3.40
2d	AFTER	2.62	7.87	BG2	2.66	6.08
2e	BG1	3.27	6.69	SBG1	2.55	4.30
2f	MQR	2.50	7.87	SBG1	2.64	4.28
3a	MQR	2.57	7.64	SBG2	2.55	3.59
3b	SAFTER	3.57	7.09	SBG2	2.26	6.14
3c	MQR	2.57	7.50	SMQR	2.50	3.87
3d	MQR	2.54	7.48	SBG2	2.59	3.92
4a	AFTER	3.00	7.84	BG1	2.48	4.44

SMAPEs para validação e teste. Para cada experimento, o método de geração de pesos selecionado foi o de melhor desempenho na validação.

Tabela 14 – Desempenhos totais para *janela mínima & pesos estáticos*

	DIESEL			*GLP*		
Experimento	*Método*	*SMAPE Validação*	*SMAPE Teste*	*Método*	*SMAPE Validação*	*SMAPE Teste*
2a	SMQI	3.36	3.31	AVG	2.48	4.79
2b	MQR	2.50	7.86	SMQI	2.46	6.76
2c	SMQI	4.20	3.99	AFTER	2.54	3.41
2d	AFTER	2.54	7.86	BG2	2.66	6.08
2e	MQI	**3.38**	**2.65**	SMQI	**2.44**	**3.36**
2f	MQR	2.50	5.18	SAFTER	2.49	3.83
3a	BG2	2.27	12.26	BG2	2.52	3.95
3b	AFTER	3.38	7.91	SMQR	2.47	3.78
3c	MQR	2.53	7.86	AVG	2.48	4.32
3d	BG2	2.20	9.65	SMQI	2.43	3.54
4a	AFTER	3.05	7.85	AVG	2.47	4.59

SMAPEs para validação e teste. Para cada experimento, o método de geração de pesos selecionado foi o de melhor desempenho na validação.

Tabela 15 – Desempenhos totais para *janela expansiva & pesos dinâmicos*

	DIESEL			*GLP*		
Experimento	*Método*	*SMAPE Validação*	*SMAPE Teste*	*Método*	*SMAPE Validação*	*SMAPE Teste*
2a	SAFTER	3.47	7.33	SBG1	2.51	4.90
2b	SAFTER	3.02	7.19	SBG1	2.59	4.95
2c	AFTER	4.20	5.40	MQI	2.46	3.92
2d	AFTER	2.92	7.85	SBG1	2.84	5.53
2e	AFTER	3.39	7.97	SBG1	2.49	4.47
2f	AFTER	2.50	7.86	SBG1	2.56	4.53
3a	BG1	3.10	7.32	SBG1	2.64	5.14
3b	SAFTER	3.57	7.07	SBG1	2.49	4.40
3c	BG2	2.99	7.38	SBG1	2.50	4.44
3d	AFTER	2.98	7.88	SBG1	2.62	4.84
4a	SAFTER	3.24	7.18	SBG1	2.52	4.71

SMAPEs para validação e teste. Para cada experimento, o método de geração de pesos selecionado foi o de melhor desempenho na validação.

Tabela 16 – Desempenhos totais para *janela expansiva & pesos estáticos*

	DIESEL			*GLP*		
Experimento	*Método*	*SMAPE Validação*	*SMAPE Teste*	*Método*	*SMAPE Validação*	*SMAPE Teste*
2a	MQR	3.61	6.83	AVG	2.48	4.79
2b	MQR	3.21	6.78	AVG	2.51	4.84
2c	MQR	4.44	5.54	MQI	2.52	3.79
2d	AFTER	2.89	7.85	AVG	2.84	5.53
2e	AFTER	3.39	7.97	AVG	2.47	4.34
2f	AFTER	2.50	7.86	SAFTER	2.48	3.94
3a	BG1	3.14	7.10	AVG	2.58	5.05
3b	MQR	3.61	6.83	MQR	2.47	4.49
3c	MQR	3.22	6.76	AVG	2.48	4.32
3d	AFTER	3.00	7.84	SAFTER	2.48	4.16
4a	MQR	3.39	6.76	AVG	2.47	4.59

SMAPEs para validação e teste. Para cada experimento, o método de geração de pesos selecionado foi o de melhor desempenho na validação.

Neste trabalho foram testadas apenas duas janelas de tempo na geração de pesos: mínima (de tamanho igual ao número de previsores sendo combinados) e expansiva. Contudo, uma análise paralela, simplificada, foi conduzida para avaliar a variabilidade dos métodos de ponderação em relação ao tamanho da janela sendo empregada. As Figuras 21 e 22 exibem os desvios-padrão médios observados nos desempenhos totais de validação (SMAPEs), quando os experimentos da Tabela 12 foram submetidos a janelas de tempo com tamanho variando do mínimo valor possível (2, 3 ou 4) até um máximo (arbitrado) de 12. Das figuras, pode-se observar que os métodos **convexos** (MQR, BG1 e AFTER) têm menor desvio-padrão médio do que os **não convexos** (MQI e BG2), indicando maior **estabilidade** dos primeiros em relação aos segundos, no que diz respeito à variações nos previsores combinados e no tamanho da janela.

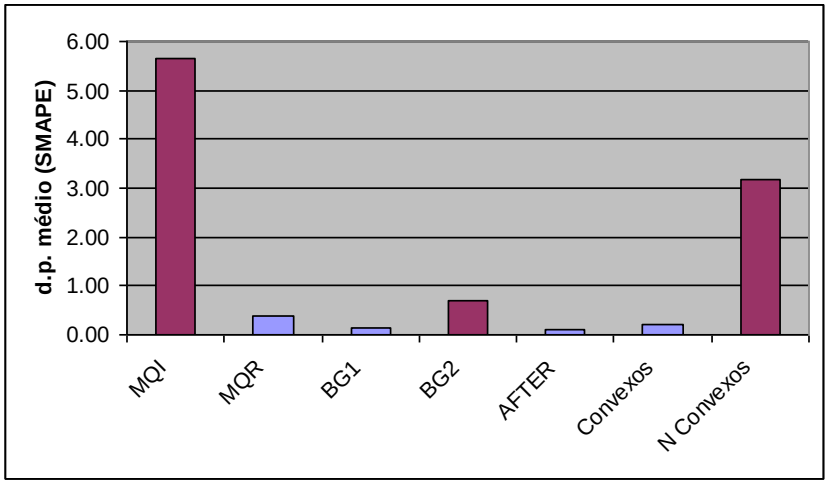

Figura 21 – Desvio padrão médio observado no conjunto de validação (DIESEL).

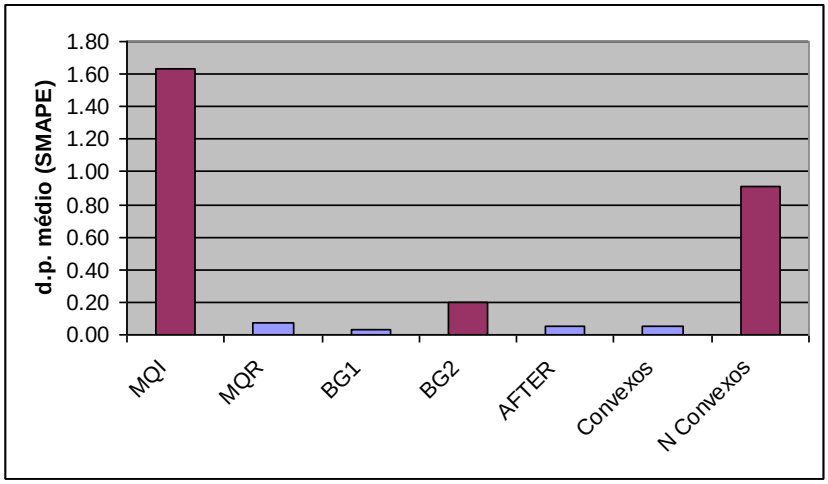

Figura 22 – Desvio padrão médio observado no conjunto de validação (GLP).

4.4.2.2. Análise individual

As Tabelas 17 e 18 e as Figuras 23 e 24 exibem a evolução dos SMAPEs ao longo do horizonte de teste, considerando os previsores individuais e a melhor combinação tradicional obtida (de menor erro composto). Na Figura 23 é interessante observar uma característica empírica das metodologias de combinação: o erro médio pode decair à medida que o horizonte de previsão aumenta, em oposição ao comportamento natural dos previsores individuais.

Tabela 17 – Desempenhos individuais e melhor combinação (DIESEL)

h	*HW*	*REG*	*DEC*	*BJ*	*COMBINAÇÃO*
1	1.61	2.56	1.01	4.46	8.20
2	3.47	3.36	4.34	3.29	4.92
3	5.11	6.01	6.85	4.80	4.56
4	5.46	6.68	7.62	4.96	3.80
5	5.92	7.40	8.27	5.43	3.72
6	6.05	7.74	8.47	5.54	3.46
7	6.32	8.22	8.81	5.85	3.51
8	6.18	8.20	8.59	5.82	3.29
9	6.02	8.24	8.42	5.70	3.01
10	5.47	7.82	7.86	5.23	3.00
11	5.48	7.80	7.83	5.20	2.82
12	5.56	7.97	7.86	5.18	2.65

Melhor combinação tradicional: *2e, geração MQI, janela mínima & pesos estáticos.*

Tabela 18 – Desempenhos individuais e melhor combinação (GLP)

h	*HW*	*REG*	*DEC*	*BJ*	*COMBINAÇÃO*
1	3.17	0.75	1.40	2.10	1.42
2	1.87	0.63	0.91	2.04	1.61
3	3.34	2.98	3.00	3.26	3.11
4	3.92	4.41	4.18	3.64	3.41
5	3.72	4.42	4.13	3.30	3.15
6	4.05	4.93	4.73	3.52	3.39
7	4.13	5.08	5.00	3.51	3.41
8	4.31	5.23	5.31	3.61	3.58
9	4.24	5.26	5.34	3.49	3.46
10	3.88	5.05	5.11	3.21	3.22
11	4.14	5.37	5.40	3.44	3.45
12	4.18	5.50	5.57	3.39	3.36

Melhor combinação tradicional: *2e, geração SMQI, janela mínima & pesos estáticos.*

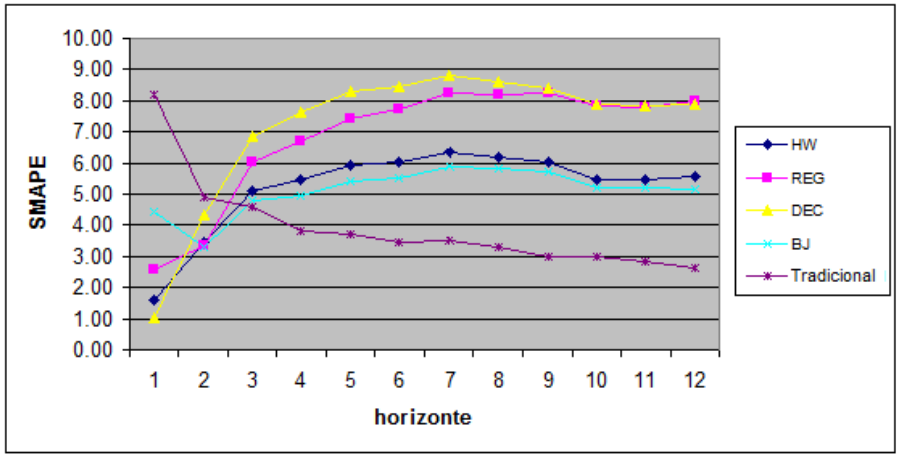

Figura 23 – Evolução dos SMAPEs fora da amostra (DIESEL).

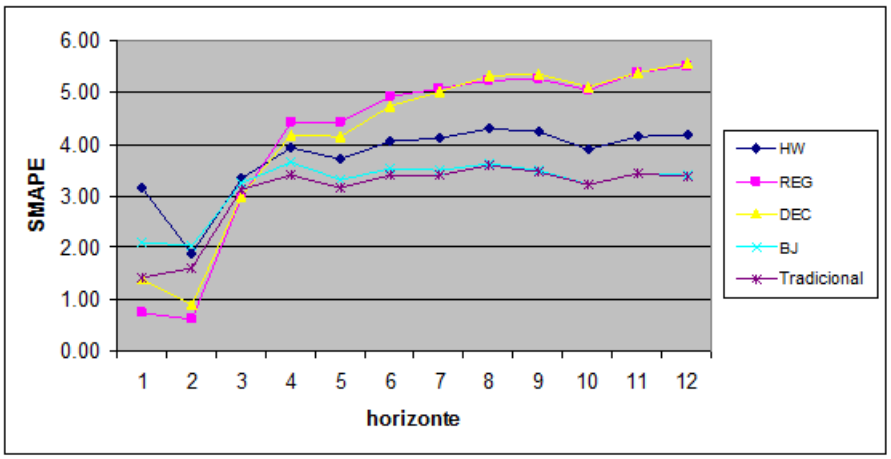

Figura 24 – Evolução dos SMAPEs fora da amostra (GLP).

As Tabelas 19 e 20 exibem as diferenças de desempenho tomadas período a período, fora da amostra, na ordem "erro do previsor individual (*benchmarking*) **menos** erro do previsor combinado". Quanto mais positiva a diferença, melhor o método à direita da comparação.

Para verificar se os desempenhos observados no teste são significativamente diferentes (a favor ou não da combinação), deve-se testar, para cada um dos métodos *benchmarking* (neste caso, os previsores individuais) a seguinte hipótese nula (H_0): a mediana das diferenças de desempenho entre o método *benchmarking* e a combinação selecionada é zero.

Tabela 19 – Diferenças de desempenho individuais/combinação (DIESEL)

h	*HW*	*REG*	*DEC*	*BJ*
1	-6.59	-5.64	-7.19	-3.74
2	-1.45	-1.56	-0.58	-1.63
3	0.55	1.44	2.29	0.23
4	1.66	2.88	3.82	1.16
5	2.19	3.68	4.55	1.70
6	2.59	4.28	5.01	2.08
7	2.81	4.71	5.30	2.34
8	2.90	4.91	5.30	2.53
9	3.01	5.23	5.41	2.69
10	2.47	4.82	4.86	2.23
11	2.66	4.98	5.00	2.38
12	2.91	5.32	5.21	2.53
MEDIANA	**2.53**	**4.50**	**4.93**	**2.15**

Tabela 20 – Diferenças de desempenho individuais/combinação (GLP)

h	*HW*	*REG*	*DEC*	*BJ*
1	1.75	-0.67	-0.02	0.68
2	0.26	-0.98	-0.70	0.43
3	0.23	-0.13	-0.11	0.15
4	0.51	1.00	0.76	0.23
5	0.56	1.27	0.98	0.15
6	0.66	1.54	1.34	0.13
7	0.71	1.66	1.58	0.09
8	0.73	1.65	1.73	0.03
9	0.78	1.80	1.88	0.03
10	0.67	1.83	1.89	-0.01
11	0.70	1.93	1.95	-0.01
12	0.82	2.13	2.21	0.03
MEDIANA	**0.68**	**1.60**	**1.46**	**0.11**

As Tabelas 21 e 22 exibem os resultados dos testes de hipóteses sugeridos na seção 4.3.2. Nas tabelas, a coluna H_0 pode assumir três valores: **0**, se a hipótese nula não for rejeitada (intervalo de confiança incluindo zero); **1** se há indicativo de que a combinação é melhor que o *benchmarking* (intervalo de confiança positivo);

-1 se há indicativo de que o *benchmarking* é melhor (intervalo de confiança negativo). Na última linha são exibidos os **saldos de rejeição da hipótese nula** (***srh0***), indicadores propostos neste trabalho e constituídos pela soma das células H_0 para cada método *benchmarking*; quanto maior este indicador, mais vezes a combinação testada foi considerada melhor.

Todas as comparações neste capítulo foram realizadas com os 3 testes propostos na seção 4.3.2. Assim, o indicador *srh0* pode assumir valores inteiros entre **-3** (a combinação é totalmente pior) e **3** (a combinação é totalmente melhor); o valor **0** indica indiferença total com o *benchmarking*. Valores do indicador entre **-1** e **1** (inclusive) constituem uma zona de indecisão, onde não se pode apontar diferença significativa entre os métodos. Com base nestas considerações, pode-se chegar às conclusões da Tabela 23; na tabela, a última linha informa o ***srh0*** **acumulado** (***srh0***+), i.e., a soma de todos os indicadores **positivos** observados.

Tabela 21 – Testes de hipótese (DIESEL)

Combinação →	*2e, geração MQI, janela mínia & pesos estáticos*				
Teste t					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	0	0.134	-0.47	3.09	Não normal
REG	1	0.012	0.79	5.06	Não normal
DEC	1	0.012	0.89	5.61	Não normal
BJ	0	0.060	-0.06	2.48	Não normal
Teste de sinais					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	1	0.039	0.55	2.89	Não normal
REG	1	0.039	1.45	4.98	Não normal
DEC	1	0.039	2.29	5.30	Não normal
BJ	1	0.039	0.24	2.53	Não normal
Teste de Wilcoxon					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	0	0.052	-0.45	2.78	Não normal
REG	0	0.052	-0.06	4.91	Não normal
DEC	1	0.043	0.86	5.16	Não normal
BJ	0	0.064	-0.53	2.39	Não normal
***srh0* (HW) = 1**	***srh0* (REG) = 2**		***srh0* (DEC) = 3**	***srh0* (BJ) = 1**	

Para todos os testes são exibidos *valor-p* (*pvalue*), limites de confiança inferior (*inf*) e superior *(sup)* para a mediana observada e status do teste de normalidade Jarque-Bera (*JB*) para a distribuição estatística das diferenças de desempenho: *normal* ou *não*.

Tabela 22 – Testes de hipótese (GLP)

Combinação →	*2e, geração SMQI, janela mínia & pesos estáticos*				
Teste t					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	1	0.000	0.46	0.94	Não normal
REG	1	0.005	0.41	1.77	Normal
DEC	1	0.002	0.52	1.73	Normal
BJ	1	0.019	0.03	0.29	Não normal
Teste de sinais					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	1	0.001	0.51	0.78	Não normal
REG	0	0.146	-0.13	1.83	Normal
DEC	0	0.146	-0.02	1.89	Normal
BJ	1	0.039	0.03	0.23	Não normal
Teste de Wilcoxon					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	1	0.001	0.48	0.80	Não normal
REG	1	0.007	0.41	1.80	Normal
DEC	1	0.007	0.48	1.81	Normal
BJ	1	0.002	0.03	0.33	Não normal
***srh0* (HW) = 3**	***srh0* (REG) = 2**	***srh0* (DEC) = 2**	***srh0* (BJ) = 3**		

Para todos os testes são exibidos *valor-p* (*pvalue*), limites de confiança inferior (*inf*) e superior *(sup)* para a mediana observada e status do teste de normalidade Jarque-Bera (*JB*) para a distribuição estatística das diferenças de desempenho: *normal* ou *não*.

Tabela 23 – Conclusões para combinação tradicional

	DIESEL		*GLP*	
Benchmarking	*srh0*	*Conclusão*	*srh0*	*Conclusão*
HW	1	Indiferente	3	A combinação é melhor
REG	2	A combinação é melhor	2	A combinação é melhor
DEC	3	A combinação é melhor	2	A combinação é melhor
BJ	1	Indiferente	3	A combinação é melhor
***srh0*+ = 17**				

Por fim, as Figuras 25 e 26 exibem a evolução[13] dos pesos de combinação selecionados, ao longo do horizonte de teste.

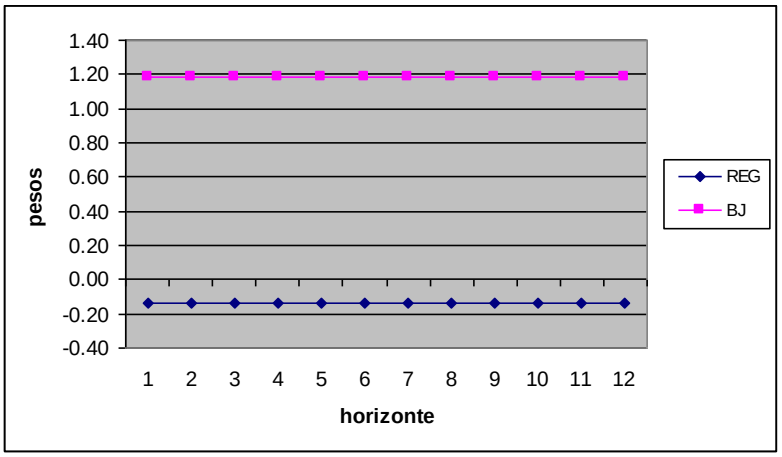

Figura 25 – Evolução dos pesos de combinação fora da amostra (DIESEL).

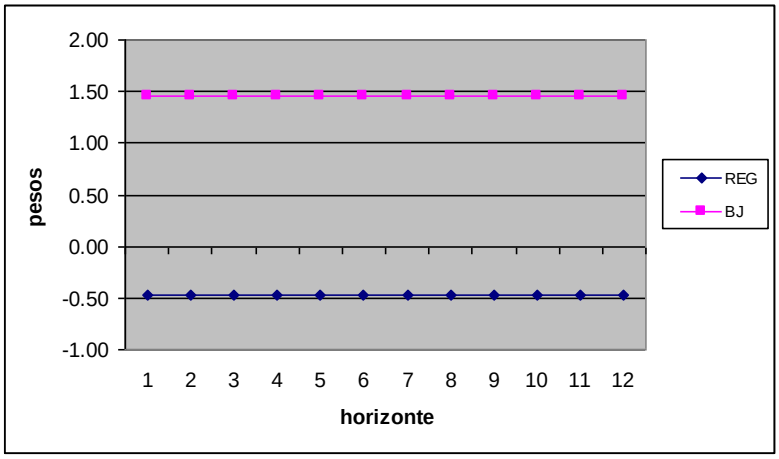

Figura 26 – Evolução dos pesos de combinação fora da amostra (GLP).

4.4.3. Combinações limiares

4.4.3.1. Experimentos

Combinações **convexa**s são não tendenciosas e limitadas pela magnitude dos previsores (seção 2.4.3.); com estas características, é intuitivo pensar que elas sejam mais estáveis (de menor variância) que as combinações **irrestritas** (não

[13] Neste caso específico não há evolução: os pesos são **estáticos**.

convexas). De fato, esta hipótese foi reforçada pela investigação específica conduzida na seção 4.4.2.1 (Figuras 21 e 22); por conta disso, e também do seu maior potencial de interpretação, incluiu-se neste trabalho uma série de experimentos focados exclusivamente na geração de pesos convexos.

Dado que o resultado de uma combinação convexa é limitado pela magnitude dos previsores combinados, propõe-se aqui um paradigma de combinação onde cada previsor original é substituído pelos seus **limites de confiança** de 95%, ou em outras palavras, cada previsor original é substituído por dois previsores **limiares**, batizados com o nome do previsor original seguido dos prefixos "+" (limite superior) ou "-" (limite inferior) (e.g. HW+, HW-). As Figuras 27 a 30 ilustram esta proposta. O limite de confiança de 95% é definido como sendo o intervalo de ± 2 desvios-padrão a partir do previsor original; considera-se, por simplificação, que o desvio-padrão é constante e vale $\sqrt{MSE}$ - raiz quadrada do erro quadrático médio de previsão, tomado dentro da amostra[14].

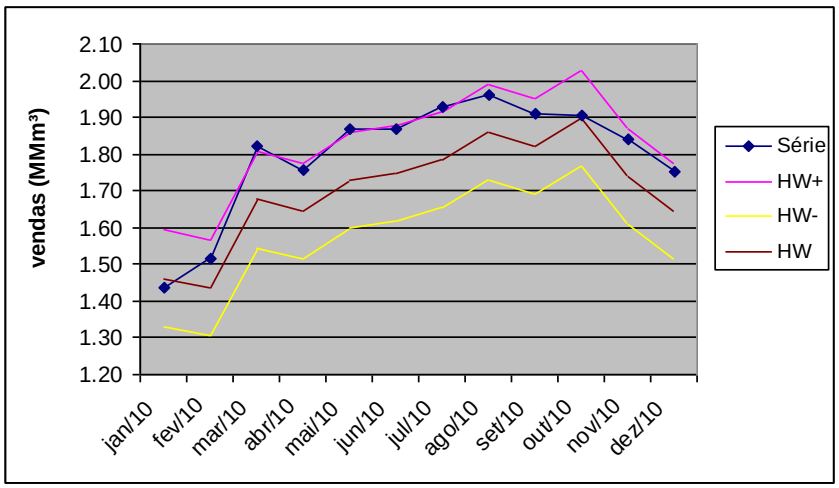

Figura 27 – Previsores limiares HW para DIESEL. HW+ e HW- são respectivamente os limites superior e inferior do previsor original (HW).

[14] Na prática, ao subtrair uma constante da previsão original, deve-se cuidar para que não ocorram previsões com valores negativos.

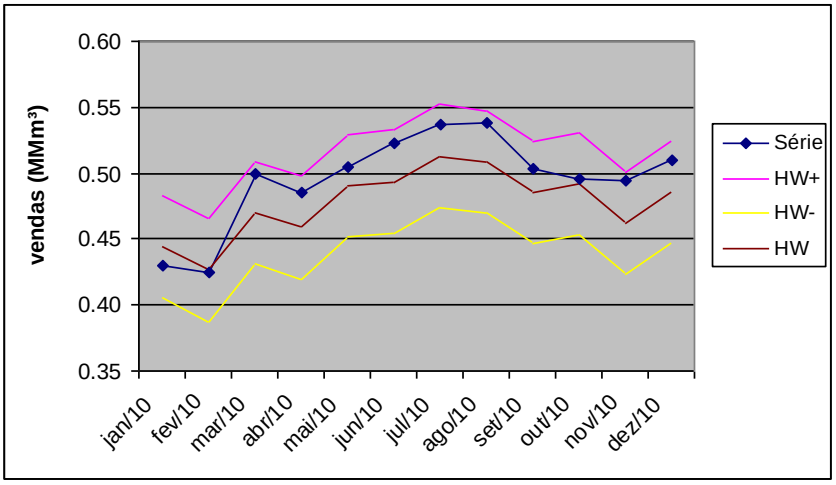

Figura 28 – Previsores limiares HW para GLP. HW+ e HW- são respectivamente os limites superior e inferior do previsor original (HW).

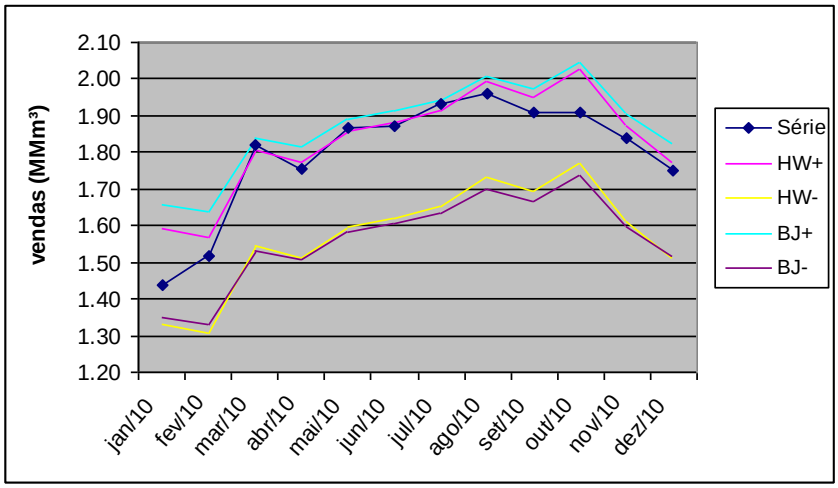

Figura 29 – Previsores limiares HW e BJ para DIESEL.

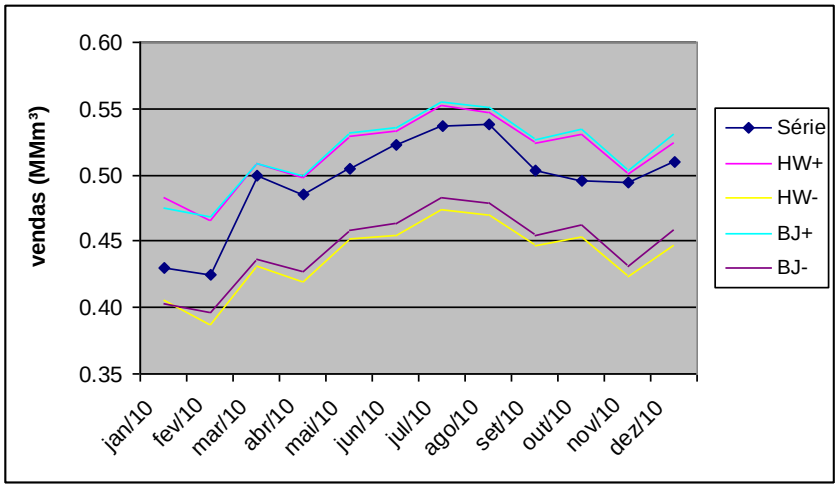

Figura 30 – Previsores limiares HW e BJ para GLP.

Ao propor o uso de previsores limiares, a ideia é privilegiar os resultados das combinações convexas, ainda que não se tenha garantia de que este tipo de combinação, mesmo com previsores limiares, seja sempre a melhor opção. É importante observar que, embora os previsores limiares originados do mesmo previsor original sejam totalmente correlacionados entre si, a **correlação de erro** entre eles pode ser, como se deseja, suficientemente baixa (seção 2.2.2).

Para avaliar o paradigma de combinação com previsores limiares, chamado neste trabalho de **combinação limiar**, foi definida a sequência de experimentos da Tabela 24, nos mesmos moldes da Tabela 12.

Tabela 24 – Experimentos de combinação limiar

Experimento	*Previsores Combinados*
1a	HW+, HW-
1b	REG+, REG-
1c	DEC+, DEC-
1d	BJ+, BJ-
22c	HW+, HW-, BJ+, BJ-

A sequência de experimentos criada privilegiou as combinações mais simples possíveis, entre previsores limiares derivados do mesmo previsor original. Assim, procura-se testar a hipótese de que uma metodologia simples, focada em combinação convexa, possa apresentar resultados comparáveis aos dos experimentos com geração de pesos irrestritos. Testou-se também uma combinação (22c) com 4 previsores limiares, derivados dos dois modelos de melhor desempenho no teste (HW e BJ, Tabelas 10 e 11).

As 5 combinações da Tabela 24 foram testadas com os métodos básicos para geração de pesos convexos (seção 2.4): média simples (AVG), mínimos quadrados restritos (MQR), Bates & Granger simples (BG1) e AFTER (AFTER). A exemplo da seção 4.4.2.1, foram testadas duas janelas de tempo para cada método de geração – mínima e expansiva; além disso, as duas formas de geração possíveis – dinâmica e estática – foram avaliadas. Deste modo, o total de experimentos foi de **80** por série.

As Tabelas 25 a 28 exibem os resultados selecionados, obtidos dentro e fora da amostra (validação e teste). Como na seção 4.4.2.1, o método de geração de

pesos foi escolhido pelo desempenho medido em um conjunto de validação, constituído pelos últimos 12 meses da amostra (imediatamente anteriores ao conjunto de teste). Os valores em negrito indicam os experimentos com menor erro composto (SMAPE validação + SMAPE teste).

Tabela 25 – Desempenhos totais para *janela mínima & pesos dinâmicos*

	DIESEL			*GLP*		
Experimento	*Método*	*SMAPE Validação*	*SMAPE Teste*	*Método*	*SMAPE Validação*	*SMAPE Teste*
1a	MQR	**2.58**	**2.32**	MQR	2.66	3.66
1b	MQR	3.54	4.40	BG1	2.62	3.27
1c	MQR	3.27	4.72	BG1	2.73	3.36
1d	MQR	3.24	2.27	MQR	3.06	2.79
22c	MQR	4.15	3.22	BG1	**2.57**	**2.93**

SMAPEs para validação e teste. Para cada experimento, o método de geração de pesos selecionado foi o de melhor desempenho na validação.

Tabela 26 – Desempenhos totais para *janela mínima & pesos estáticos*

	DIESEL			*GLP*		
Experimento	*Método*	*SMAPE Validação*	*SMAPE Teste*	*Método*	*SMAPE Validação*	*SMAPE Teste*
1a	MQR	3.37	2.34	AVG	2.67	4.18
1b	AVG	3.39	7.97	BG1	2.62	3.27
1c	AVG	2.50	7.86	BG1	2.73	3.36
1d	AFTER	4.19	3.10	AVG	2.54	3.39
22c	BG1	3.82	2.09	BG1	2.50	3.09

SMAPEs para validação e teste. Para cada experimento, o método de geração de pesos selecionado foi o de melhor desempenho na validação.

Tabela 27 – Desempenhos totais para *janela expansiva & pesos dinâmicos*

	DIESEL			*GLP*		
Experimento	*Método*	*SMAPE Validação*	*SMAPE Teste*	*Método*	*SMAPE Validação*	*SMAPE Teste*
1a	MQR	4.82	5.21	BG1	2.61	3.70
1b	AFTER	3.16	9.23	MQR	2.83	5.59
1c	MQR	2.50	7.91	MQR	3.02	5.65
1d	MQR	4.63	5.21	MQR	2.52	3.85
22c	MQR	4.55	5.14	MQR	2.45	3.80

SMAPEs para validação e teste. Para cada experimento, o método de geração de pesos selecionado foi o de melhor desempenho na validação.

Tabela 28 – Desempenhos totais para *janela expansiva & pesos estáticos*

	DIESEL			*GLP*		
Experimento	*Método*	*SMAPE Validação*	*SMAPE Teste*	*Método*	*SMAPE Validação*	*SMAPE Teste*
1a	AVG	4.44	5.56	BG1	2.62	4.36
1b	AVG	3.39	7.97	AVG	2.74	5.50
1c	AVG	2.50	7.86	AVG	2.93	5.57
1d	AVG	4.51	5.18	MQR	2.50	3.53
22c	AVG	4.48	5.37	BG1	2.52	3.96

SMAPEs para validação e teste. Para cada experimento, o método de geração de pesos selecionado foi o de melhor desempenho na validação.

4.4.3.2. Análise individual

As Tabelas 29 e 30 e as Figuras 31 e 32 exibem a evolução dos SMAPEs ao longo do horizonte de teste, considerando os previsores individuais e a melhor combinação limiar obtida (de menor erro composto). Nas figuras, percebe-se decaimento do erro médio da combinação à medida que o horizonte de previsão aumenta.

Tabela 29 – Desempenhos individuais e melhor combinação (DIESEL)

h	*HW*	*REG*	*DEC*	*BJ*	*COMBINAÇÃO*
1	1.61	2.56	1.01	4.46	7.19
2	3.47	3.36	4.34	3.29	4.25
3	5.11	6.01	6.85	4.80	3.52
4	5.46	6.68	7.62	4.96	2.74
5	5.92	7.40	8.27	5.43	2.59
6	6.05	7.74	8.47	5.54	2.26
7	6.32	8.22	8.81	5.85	2.20
8	6.18	8.20	8.59	5.82	1.99
9	6.02	8.24	8.42	5.70	1.88
10	5.47	7.82	7.86	5.23	2.09
11	5.48	7.80	7.83	5.20	2.11
12	5.56	7.97	7.86	5.18	2.32

Melhor combinação limiar: *1a, geração MQR, janela mínima & pesos dinâmicos.*

Tabela 30 – Desempenhos individuais e melhor combinação (GLP)

h	*HW*	*REG*	*DEC*	*BJ*	*COMBINAÇÃO*
1	3.17	0.75	1.40	2.10	3.98
2	1.87	0.63	0.91	2.04	3.38
3	3.34	2.98	3.00	3.26	3.49
4	3.92	4.41	4.18	3.64	3.22
5	3.72	4.42	4.13	3.30	2.70
6	4.05	4.93	4.73	3.52	2.53
7	4.13	5.08	5.00	3.51	2.34
8	4.31	5.23	5.31	3.61	2.44
9	4.24	5.26	5.34	3.49	2.39
10	3.88	5.05	5.11	3.21	2.16
11	4.14	5.37	5.40	3.44	2.67
12	4.18	5.50	5.57	3.39	2.93

Melhor combinação limiar: *22c, geração BG1, janela mínima & pesos dinâmicos.*

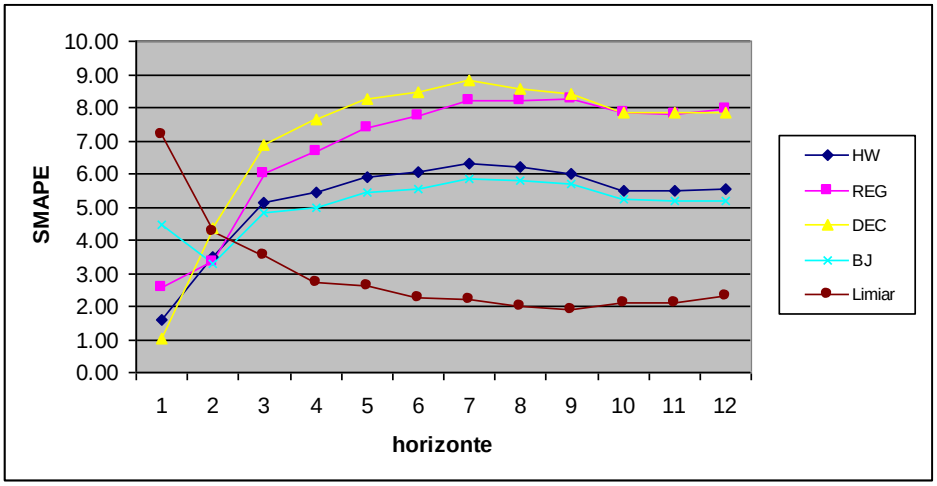

Figura 31 – Evolução dos SMAPEs fora da amostra (DIESEL).

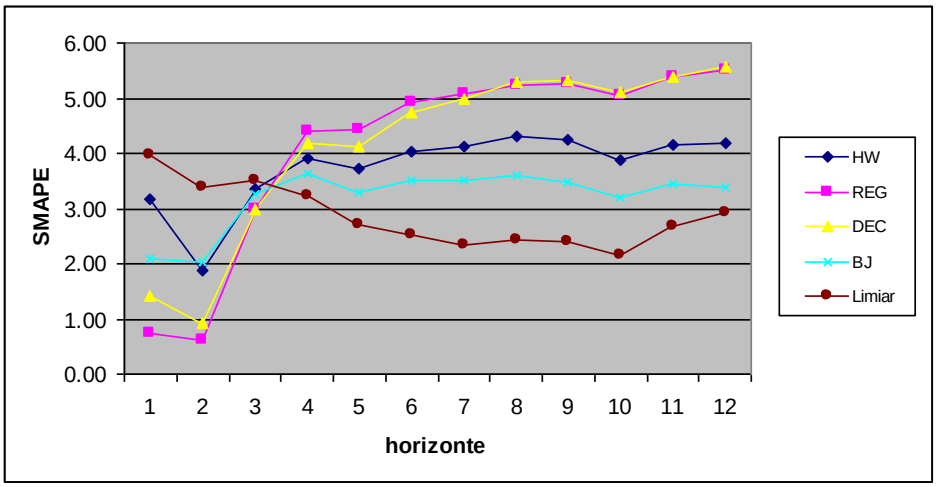

Figura 32 – Evolução dos SMAPEs fora da amostra (GLP).

As Tabelas 31 e 32 exibem as diferenças de desempenho tomadas período a período, fora da amostra, na ordem "erro do previsor individual (*benchmarking*) **menos** erro do previsor combinado". Calculadas as diferenças, todos os testes de hipótese sugeridos na seção 4.3.2 foram executados, com o objetivo de verificar se as medianas das mesmas são diferentes de zero (i.e., se há diferença significativa entre os métodos comparados). As Tabelas 33 e 34 exibem os resultados.

Tabela 31 – Diferenças de desempenho individuais/combinação (DIESEL)

h	*HW*	*REG*	*DEC*	*BJ*
1	-5.58	-4.63	-6.17	-2.73
2	-0.78	-0.89	0.09	-0.96
3	1.59	2.48	3.33	1.27
4	2.72	3.94	4.88	2.22
5	3.32	4.81	5.68	2.83
6	3.79	5.48	6.21	3.28
7	4.11	6.02	6.61	3.65
8	4.19	6.20	6.59	3.82
9	4.14	6.36	6.55	3.82
10	3.38	5.73	5.77	3.14
11	3.37	5.69	5.72	3.09
12	3.24	5.65	5.54	2.85
MEDIANA	**3.35**	**5.57**	**5.70**	**2.97**

Tabela 32 – Diferenças de desempenho individuais/combinação (GLP)

h	*HW*	*REG*	*DEC*	*BJ*
1	-0.81	-3.23	-2.58	-1.88
2	-1.51	-2.75	-2.47	-1.34
3	-0.15	-0.52	-0.50	-0.23
4	0.70	1.19	0.96	0.42
5	1.02	1.72	1.43	0.61
6	1.52	2.40	2.21	0.99
7	1.78	2.74	2.65	1.16
8	1.87	2.79	2.87	1.17
9	1.85	2.87	2.95	1.10
10	1.72	2.88	2.95	1.04
11	1.47	2.70	2.73	0.77
12	1.25	2.56	2.64	0.46
MEDIANA	**1.36**	**2.48**	**2.42**	**0.69**

Tabela 33 – Testes de hipótese (DIESEL)

Combinação →	*1a, geração MQR, janela mínia & pesos dinâmicos*				
Teste t					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	1	0.018	0.49	4.10	Não normal
REG	1	0.002	1.75	6.06	Não normal
DEC	1	0.003	1.84	6.62	Não normal
BJ	1	0.004	0.89	3.50	Não normal
Teste de sinais					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	1	0.039	1.59	4.12	Não normal
REG	1	0.039	2.49	6.02	Não normal
DEC	1	0.006	3.33	6.54	Não normal
BJ	1	0.039	1.28	3.65	Não normal
Teste de Wilcoxon					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	1	0.043	0.41	3.76	Não normal
REG	1	0.005	0.87	5.88	Não normal
DEC	1	0.012	1.71	6.16	Não normal
BJ	1	0.005	0.55	3.40	Não normal
***srh0* (HW) = 3**	***srh0* (REG) = 3**	***srh0* (DEC) = 3**	***srh0* (BJ) = 3**		

Para todos os testes são exibidos *valor-p* (*pvalue*), limites de confiança inferior (*inf*) e superior *(sup)* para a mediana observada e status do teste de normalidade Jarque-Bera (*JB*) para a distribuição estatística das diferenças de desempenho: *normal* ou *não*.

Tabela 34 – Testes de hipótese (GLP)

Combinação →	*22c, geração BG1, janela mínia & pesos dinâmicos*				
Teste t					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	1	0.020	0.18	1.61	Normal
REG	0	0.070	-0.13	2.70	Não normal
DEC	1	0.050	0.01	2.63	Normal
BJ	0	0.250	-0.29	1.00	Não normal
Teste de sinais					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	0	0.150	-0.15	1.79	Normal
REG	0	0.150	-0.51	2.79	Não normal
DEC	0	0.150	-0.49	2.87	Normal
BJ	0	0.150	-0.23	1.10	Não normal
Teste de Wilcoxon					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	1	0.030	0.10	1.66	Normal
REG	0	0.180	-0.24	2.73	Não normal
DEC	1	0.030	0.04	2.76	Normal
BJ	0	0.270	-0.39	1.02	Não normal
***srh0* (HW) = 2**	***srh0* (REG) = 0**	***srh0* (DEC) = 2**	***srh0* (BJ) = 0**		

Para todos os testes são exibidos *valor-p* (*pvalue*), limites de confiança inferior (*inf*) e superior *(sup)* para a mediana observada e status do teste de normalidade Jarque-Bera (*JB*) para a distribuição estatística das diferenças de desempenho: *normal* ou *não*.

Com base nos valores observados (Tabelas 33 e 34) para o indicador *srh0* (seção 4.4.2.2), pode-se chegar às conclusões da Tabela 35.

Tabela 35 – Conclusões para combinação limiar

	DIESEL		*GLP*	
Benchmarking	*srh0*	*Conclusão*	*srh0*	*Conclusão*
HW	3	A combinação é melhor	2	A combinação é melhor
REG	3	A combinação é melhor	0	Indiferente
DEC	3	A combinação é melhor	2	A combinação é melhor
BJ	3	A combinação é melhor	0	Indiferente
***srh0*+ = 16**				

As Figuras 33 e 34 exibem a evolução dos pesos de combinação selecionados, ao longo do horizonte de teste.

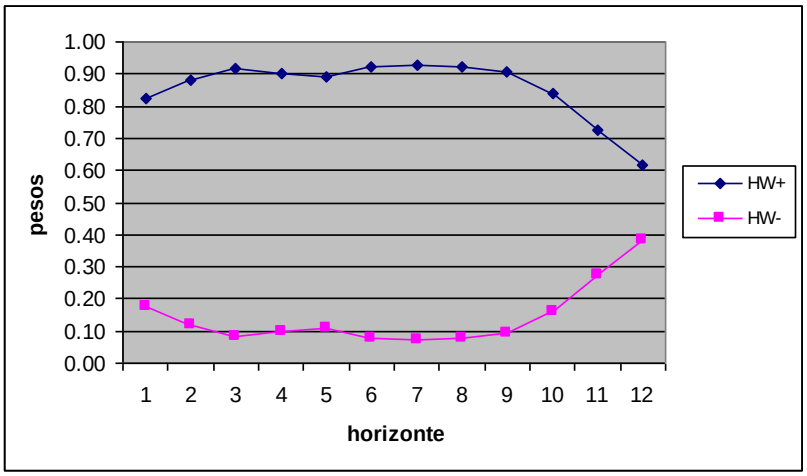

Figura 33 – Evolução dos pesos de combinação fora da amostra (DIESEL).

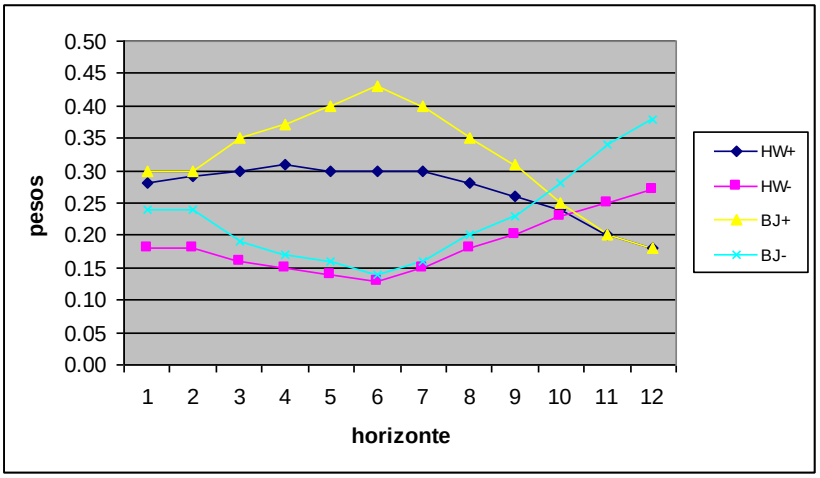

Figura 34 – Evolução dos pesos de combinação fora da amostra (GLP).

4.4.4. Combinações NEW

4.4.4.1. Experimentos

Para reduzir o escopo da análise de maneira conveniente, todos os experimentos NEW neste trabalho estiveram focados na geração de pesos convexos, de acordo com a estratégia de **previsores limiares** sugerida na seção 4.4.3.1. O embasamento teórico para esta decisão foi o fato dos modelos de geração convexa tenderem a ser mais estáveis do que os de geração irrestrita,

simplificando, do ponto de vista numérico, o treinamento das redes neurais envolvidas (seção 3.8). Além disso, este tipo de combinação pode apresentar bons resultados (seção 4.4.3.2). Com base nestas considerações, foi definida a sequência de experimentos da Tabela 36, organizada de acordo com os previsores (limiares) sendo combinados, o método de geração de **pesos históricos** e a métrica de **desempenho aproximado** (capítulo 3). Assim como na seção 4.4.3.1, os experimentos com 4 previsores limiares (prefixo "new2c") foram derivados dos modelos que apresentaram os melhores desempenhos no teste (HW e BJ, Tabelas 10 e 11).

Tabela 36 – Experimentos de combinação NEW

Experimento	*Previsores*	*Geração*	*Desempenho*
new1a_after_mad	HW+, HW-	AFTER	MAD
new1a_after_mse	HW+, HW-	AFTER	MSE
new1a_bg1_mad	HW+, HW-	BG1	MAD
new1a_bg1_mse	HW+, HW-	BG1	MSE
new1a_mqr_mad	HW+, HW-	MQR	MAD
new1a_mqr_mse	HW+, HW-	MQR	MSE
new1a_pool_mad	HW+, HW-	POOL	MAD
new1a_pool_mse	HW+, HW-	POOL	MSE
new2c_after_mad	HW+, HW-, BJ+, BJ-	AFTER	MAD
new2c_after_mse	HW+, HW-, BJ+, BJ-	AFTER	MSE
new2c_bg1_mad	HW+, HW-, BJ+, BJ-	BG1	MAD
new2c_bg1_mse	HW+, HW-, BJ+, BJ-	BG1	MSE
new2c_mqr_mad	HW+, HW-, BJ+, BJ-	MQR	MAD
new2c_mqr_mse	HW+, HW-, BJ+, BJ-	MQR	MSE
new2c_pool_mad	HW+, HW-, BJ+, BJ-	POOL	MAD
new2c_pool_mse	HW+, HW-, BJ+, BJ-	POOL	MSE

Além dos métodos básicos para geração de pesos convexos – AFTER, BG1 e MQR – foi introduzido um quarto, chamado de POOL. O método POOL refere-se não a um método isolado, mas ao esquema **híbrido** denominado neste trabalho de *pool* de métodos (seção 3.4); neste esquema, considera-se, para cada bloco de previsões no conjunto de treinamento, o melhor entre os métodos AFTER, BG1, MQR ou AVG (média simples).

Em termos práticos, o uso do sistema NEW está sujeito a cinco hiperparâmetros:

1. Série de referência;
2. Métrica de desempenho aproximado (MAD ou MSE);
3. Método de geração de pesos históricos;
4. Tamanho da janela de tempo;
5. Tipo de normalização (padrão ou soma-1).

Especificamente com relação às séries de referências, foram testados quatro tipos diferentes, de acordo com as metodologias sugeridas na seção 3.3: (i) repetição do último período (SALY1), (ii) repetição do último período com crescimento sazonal mais recente (SALY2), (iii) repetição do último período com crescimento sazonal médio (SALY3) e (iv) decomposição clássica (DEC). A Tabela 37 exibe os desempenhos totais destas referências.

Tabela 37 – Desempenhos totais (referências)

	DIESEL		*GLP*	
Referência	*SMAPE Amostra*	*SMAPE Teste*	*SMAPE Amostra*	*SMAPE Teste*
SALY1	5.09	9.97	3.55	4.37
SALY2	6.38	12.37	5.44	6.94
SALY3	4.62	7.54	3.49	5.17
DEC	3.27	7.86	2.96	5.57

As 16 combinações da Tabela 36 foram testadas com as quatro referências disponíveis, sendo a melhor referência selecionada pelo menor valor do seguinte critério: erro da previsão combinada no conjunto de validação **somado** à **correlação conjunta** dos erros, tomada dentro da amostra; este critério segue a definição de **critério misto** (seção 3.3). Na prática, as parcelas do critério devem ser postas na mesma escala, para evitar distorções. Como será visto mais adiante, isso é feito forçando-se com que cada parcela esteja no intervalo [0,1], dividindo-se tanto o vetor de desempenhos (parcela 1) quanto o vetor de correlações conjuntas (parcela 2) pelas suas respectivas somas.

As Tabelas 38 e 39 exibem as **correlações de erro** (seção 2.2.2) entre as referências e os previsores utilizados (HW e BJ) – deve-se observar que a correlação de erro entre uma referência e um previsor limiar é idêntica à correlação de erro entre esta referência e o previsor original correspondente.

Tabela 38 – Correlações de erro dentro da amostra (DIESEL)

	SALY1	*SALY2*	*SALY3*	DEC	HW	BJ
SALY1	**1.00**					
SALY2	0.80	**1.00**				
SALY3	0.99	0.81	**1.00**			
DEC	0.65	0.19	0.63	**1.00**		
HW	0.81	0.55	0.80	0.72	**1.00**	
BJ	0.70	0.43	0.67	0.65	0.93	**1.00**

Tabela 39 – Correlações de erro dentro da amostra (GLP)

	SALY1	*SALY2*	*SALY3*	DEC	HW	BJ
SALY1	**1.00**					
SALY2	0.81	**1.00**				
SALY3	0.99	0.81	**1.00**			
DEC	0.63	0.21	0.62	**1.00**		
HW	0.87	0.60	0.86	0.68	**1.00**	
BJ	0.67	0.42	0.66	0.65	0.85	**1.00**

Neste trabalho, correlação conjunta é um valor único para medir a relação linear entre uma dada referência e o conjunto de todos os previsores; ela é aproximada pela média das *N* correlações de erro referência/previsor observadas. Por exemplo, com base nos dados da Tabela 38 (DIESEL), as correlações conjuntas das referências SALY1, SALY2, SALY3 e DEC em relação ao par de previsores HW/BJ seriam, respectivamente, 0.76, 0.49, 0.74 e 0.69. Como mencionado antes, para deixar estes valores no intervalo [0,1] – e assim evitar distorções na aplicação do critério misto – basta dividi-los por sua soma (2.67), obtendo, respectivamente, 0.28, 0.18, 0.28 e 0.26.

Nos experimentos da Tabela 36, além de variar a referência utilizada, testaram-se os dois tipos de normalização possíveis (padrão e soma-1) e as duas

janelas de tempo usadas anteriormente (seções 4.4.2.1 e 4.4.3.1): mínima e expansiva. Deste modo, o total de experimentos foi de **256** por série. Com relação à escolha da melhor arquitetura de rede neural em um dado experimento, selecionou-se aquela com menor erro de **previsão combinada** (no conjunto de validação) (seção 3.9). As Tabelas 40 a 43 exibem os resultados selecionados, obtidos nos conjuntos de validação e teste; os valores em negrito indicam os experimentos com menor erro composto (SMAPE validação + SMAPE teste).

Tabela 40 – Desempenhos totais para *janela mínima & normalização padrão*

	DIESEL			*GLP*		
Experimento	*Referência*	*SMAPE Validação*	*SMAPE Teste*	*Referência*	*SMAPE Validação*	*SMAPE Teste*
new1a_after_mad	SALY2	3.99	3.77	SALY2	3.64	3.20
new1a_after_mse	SALY2	4.32	2.98	SALY2	2.89	3.97
new1a_bg1_mad	SALY2	4.06	3.10	SALY2	3.05	3.35
new1a_bg1_mse	SALY2	4.05	3.47	SALY2	3.10	3.82
new1a_mqr_mad	SALY2	3.91	2.65	SALY2	2.82	3.39
new1a_mqr_mse	SALY2	4.26	2.41	SALY2	3.10	3.98
new1a_pool_mad	SALY2	3.90	4.02	SALY2	2.98	3.76
new1a_pool_mse	SALY2	4.13	3.56	SALY2	2.92	3.41
new2c_after_mad	SALY2	4.91	4.78	SALY2	2.96	3.72
new2c_after_mse	SALY2	4.39	3.41	SALY2	4.10	3.94
new2c_bg1_mad	SALY2	3.94	4.37	SALY2	3.01	3.46
new2c_bg1_mse	SALY2	5.00	3.26	SALY2	3.32	3.63
new2c_mqr_mad	SALY2	3.81	6.26	SALY2	2.94	4.25
new2c_mqr_mse	SALY2	3.86	3.82	SALY2	3.35	4.02
new2c_pool_mad	SALY2	3.89	3.74	SALY2	2.93	3.79
new2c_pool_mse	SALY2	4.09	4.25	SALY2	3.13	3.90

Para cada experimento, a série de referência foi selecionada pelo menor critério misto: *erro na validação + correlação conjunta*.

Tabela 41 – Desempenhos totais para *janela mínima & normalização soma-1*

	DIESEL			*GLP*		
Experimento	*Referência*	*SMAPE Validação*	*SMAPE Teste*	*Referência*	*SMAPE Validação*	*SMAPE Teste*
new1a_after_mad	SALY2	3.73	3.10	DEC	2.53	4.15
new1a_after_mse	SALY2	3.91	2.54	DEC	2.61	3.31
new1a_bg1_mad	SALY2	3.80	4.87	SALY2	2.93	3.76
new1a_bg1_mse	SALY2	3.76	3.97	SALY2	2.93	3.59
new1a_mqr_mad	SALY2	4.01	4.21	SALY2	2.96	3.82
new1a_mqr_mse	SALY2	3.86	4.32	SALY2	2.86	3.97
new1a_pool_mad	SALY2	4.08	4.28	SALY2	2.83	3.57
new1a_pool_mse	SALY2	3.83	4.17	SALY2	2.94	3.79
new2c_after_mad	SALY2	3.60	3.50	SALY2	2.64	4.88
new2c_after_mse	SALY2	4.33	5.88	SALY2	3.20	4.36
new2c_bg1_mad	SALY2	3.01	3.26	SALY2	3.06	4.54
new2c_bg1_mse	SALY2	4.08	3.59	SALY2	2.75	4.24
new2c_mqr_mad	SALY2	3.83	3.37	SALY2	2.90	4.10
new2c_mqr_mse	SALY2	3.53	3.97	SALY2	2.61	3.33
new2c_pool_mad	SALY2	3.49	3.95	SALY2	2.47	4.44
new2c_pool_mse	SALY2	3.47	3.52	SALY2	2.87	4.45

Para cada experimento, a série de referência foi selecionada pelo menor critério misto:
erro na validação + correlação conjunta.

Tabela 42 – Desempenhos totais para *janela expansiva & normalização padrão*

	DIESEL			*GLP*		
Experimento	*Referência*	*SMAPE Validação*	*SMAPE Teste*	*Referência*	*SMAPE Validação*	*SMAPE Teste*
new1a_after_mad	SALY2	3.71	4.33	SALY2	3.17	3.58
new1a_after_mse	SALY1	2.97	8.40	SALY2	2.98	2.31
new1a_bg1_mad	SALY2	3.39	8.53	SALY2	2.92	2.88
new1a_bg1_mse	SALY2	4.53	6.63	SALY2	3.15	3.37
new1a_mqr_mad	SALY2	3.35	3.04	SALY2	2.94	3.11
new1a_mqr_mse	SALY2	**3.82**	**2.07**	SALY2	3.00	2.64
new1a_pool_mad	SALY2	3.34	4.86	SALY2	2.91	4.21
new1a_pool_mse	SALY2	3.96	5.32	SALY2	2.92	2.77
new2c_after_mad	SALY2	4.09	3.71	SALY2	3.48	7.46
new2c_after_mse	SALY2	3.84	2.64	SALY2	3.19	5.02
new2c_bg1_mad	SALY2	3.34	2.78	SALY2	2.72	4.92
new2c_bg1_mse	SALY2	3.33	7.92	SALY2	2.77	3.68
new2c_mqr_mad	SALY2	3.34	4.80	SALY2	2.78	3.37
new2c_mqr_mse	SALY2	3.56	4.90	SALY2	2.73	2.77
new2c_pool_mad	SALY2	3.28	2.93	SALY2	3.91	3.74
new2c_pool_mse	SALY2	3.43	3.30	SALY2	3.87	2.94

Para cada experimento, a série de referência foi selecionada pelo menor critério misto:

erro na validação + correlação conjunta.

Tabela 43 – Desempenhos totais para *janela expansiva & normalização soma-1*

	DIESEL			*GLP*		
Experimento	*Referência*	*SMAPE Validação*	*SMAPE Teste*	*Referência*	*SMAPE Validação*	*SMAPE Teste*
new1a_after_mad	SALY2	3.92	4.18	SALY2	2.88	3.81
new1a_after_mse	SALY2	3.93	3.35	SALY2	2.91	2.55
new1a_bg1_mad	SALY2	4.03	2.61	SALY2	2.86	3.28
new1a_bg1_mse	SALY2	4.16	3.35	SALY2	2.87	3.31
new1a_mqr_mad	SALY2	3.84	4.62	SALY2	2.89	4.06
new1a_mqr_mse	SALY2	3.98	5.03	SALY2	2.86	3.39
new1a_pool_mad	SALY2	3.72	4.10	SALY2	2.86	3.85
new1a_pool_mse	SALY2	4.01	4.74	SALY2	2.88	3.73
new2c_after_mad	SALY2	4.95	7.17	SALY2	3.27	3.64
new2c_after_mse	SALY2	4.65	3.70	SALY2	2.76	3.48
new2c_bg1_mad	SALY2	3.97	3.76	SALY2	2.74	3.35
new2c_bg1_mse	SALY2	4.10	4.73	SALY2	2.79	3.61
new2c_mqr_mad	SALY2	3.58	5.13	SALY2	2.80	3.43
new2c_mqr_mse	SALY2	4.04	5.33	SALY2	2.71	3.57
new2c_pool_mad	SALY2	4.03	4.60	SALY2	**2.85**	**2.32**
new2c_pool_mse	SALY2	3.97	3.03	SALY2	2.83	2.55

Para cada experimento, a série de referência foi selecionada pelo menor critério misto:
erro na validação + correlação conjunta.

4.4.4.2. Análise individual

As Tabelas 44 e 45 e as Figuras 35 e 36 exibem a evolução dos SMAPEs ao longo do horizonte de teste, considerando os previsores individuais e a melhor combinação NEW obtida (de menor erro composto). Nas figuras, percebe-se o decaimento do erro médio da combinação à medida que o horizonte de previsão aumenta.

Tabela 44 – Desempenhos individuais e melhor combinação (DIESEL)

h	*HW*	*REG*	*DEC*	*BJ*	*COMBINAÇÃO*
1	1.61	2.56	1.01	4.46	5.76
2	3.47	3.36	4.34	3.29	3.31
3	5.11	6.01	6.85	4.80	2.94
4	5.46	6.68	7.62	4.96	2.41
5	5.92	7.40	8.27	5.43	2.51
6	6.05	7.74	8.47	5.54	2.48
7	6.32	8.22	8.81	5.85	2.50
8	6.18	8.20	8.59	5.82	2.19
9	6.02	8.24	8.42	5.70	2.00
10	5.47	7.82	7.86	5.23	2.27
11	5.48	7.80	7.83	5.20	2.14
12	5.56	7.97	7.86	5.18	2.07

Melhor combinação NEW: *new1a_mqr_mse, SALY2, janela expansiva & normalização padrão.*

Tabela 45 – Desempenhos individuais e melhor combinação (GLP)

h	*HW*	*REG*	*DEC*	*BJ*	*COMBINAÇÃO*
1	3.17	0.75	1.40	2.10	0.91
2	1.87	0.63	0.91	2.04	1.88
3	3.34	2.98	3.00	3.26	3.20
4	3.92	4.41	4.18	3.64	3.38
5	3.72	4.42	4.13	3.30	3.11
6	4.05	4.93	4.73	3.52	3.04
7	4.13	5.08	5.00	3.51	2.70
8	4.31	5.23	5.31	3.61	2.51
9	4.24	5.26	5.34	3.49	2.28
10	3.88	5.05	5.11	3.21	2.36
11	4.14	5.37	5.40	3.44	2.46
12	4.18	5.50	5.57	3.39	2.32

Melhor combinação NEW: *new2c_pool_mad, SALY2, janela expansiva & normalização soma-1.*

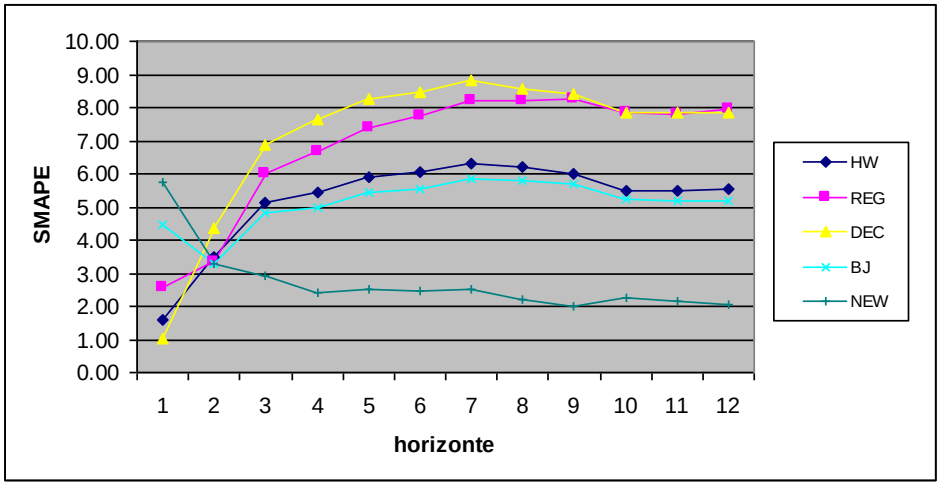

Figura 35 – Evolução dos SMAPEs fora da amostra (DIESEL).

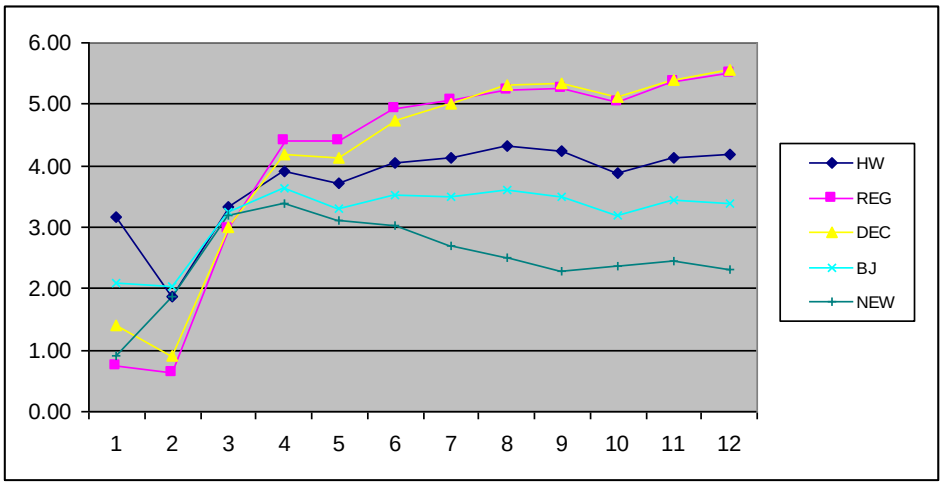

Figura 36 – Evolução dos SMAPEs fora da amostra (GLP).

As Tabelas 46 e 47 exibem as diferenças de desempenho tomadas período a período, fora da amostra, na ordem "erro do previsor individual (*benchmarking*) **menos** erro do previsor combinado". Como em seções anteriores, todos os testes de hipótese sugeridos para comparação de desempenhos foram executados; as Tabelas 48 e 49 exibem os resultados.

Tabela 46 – Diferenças de desempenho individuais/combinação (DIESEL)

h	*HW*	*REG*	*DEC*	*BJ*
1	-4.15	-3.20	-4.75	-1.30
2	0.17	0.05	1.03	-0.01
3	2.17	3.07	3.92	1.86
4	3.04	4.26	5.21	2.54
5	3.41	4.89	5.76	2.92
6	3.57	5.26	5.99	3.06
7	3.82	5.72	6.31	3.35
8	3.99	6.00	6.39	3.62
9	4.02	6.24	6.43	3.71
10	3.20	5.55	5.59	2.96
11	3.34	5.66	5.69	3.06
12	3.49	5.90	5.79	3.10
MEDIANA	**3.37**	**5.41**	**5.73**	**3.01**

Tabela 47 – Diferenças de desempenho individuais/combinação (GLP)

h	*HW*	*REG*	*DEC*	*BJ*
1	2.26	-0.16	0.49	1.19
2	-0.01	-1.25	-0.97	0.16
3	0.14	-0.22	-0.20	0.06
4	0.55	1.03	0.80	0.27
5	0.60	1.30	1.02	0.19
6	1.01	1.90	1.70	0.49
7	1.43	2.38	2.30	0.81
8	1.80	2.72	2.80	1.10
9	1.96	2.98	3.06	1.21
10	1.52	2.68	2.75	0.84
11	1.68	2.91	2.93	0.98
12	1.86	3.17	3.24	1.07
MEDIANA	**1.48**	**2.14**	**2.00**	**0.83**

Tabela 48 – Testes de hipótese (DIESEL)

Combinação →	*new1a_mqr_mse, SALY2, janela expansiva & normalização padrão*				
Teste t					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	1	0.004	1.02	4.00	Não normal
REG	1	0.000	2.29	5.95	Não normal
DEC	1	0.001	2.37	6.52	Não normal
BJ	1	0.000	1.43	3.38	Não normal
Teste de sinais					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	1	0.006	2.17	3.82	Não normal
REG	1	0.006	3.07	5.90	Não normal
DEC	1	0.006	3.91	6.31	Não normal
BJ	1	0.039	1.86	3.35	Não normal
Teste de Wilcoxon					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	1	0.034	1.17	3.68	Não normal
REG	1	0.002	1.56	5.78	Não normal
DEC	1	0.002	2.47	6.06	Não normal
BJ	1	0.002	1.20	3.31	Não normal
***srh0* (HW) = 3**	***srh0* (REG) = 3**	***srh0* (DEC) = 3**	***srh0* (BJ) = 3**		

Para todos os testes são exibidos *valor-p* (*pvalue*), limites de confiança inferior (*inf*) e superior *(sup)* para a mediana observada e status do teste de normalidade Jarque-Bera (*JB*) para a distribuição estatística das diferenças de desempenho: *normal* ou *não*.

Tabela 49 – Testes de hipótese (GLP)

Combinação→	*new2c_pool_mad, SALY2, janela expansiva & normalização soma 1*				
Teste t					
Benchmarking	*H_0*	*pvalue*	*inf*	*sup*	*JB*
HW	1	0.000	0.75	1.71	Normal
REG	1	0.003	0.68	2.56	Normal
DEC	1	0.002	0.76	2.56	Normal
BJ	1	0.000	0.42	0.98	Normal
Teste de sinais					
Benchmarking	*H_0*	*pvalue*	*inf*	*sup*	*JB*
HW	1	0.006	0.54	1.86	Normal
REG	0	0.146	-0.16	2.91	Normal
DEC	1	0.039	0.49	2.94	Normal
BJ	1	0.001	0.19	1.10	Normal
Teste de Wilcoxon					
Benchmarking	*H_0*	*pvalue*	*inf*	*sup*	*JB*
HW	1	0.001	0.76	1.80	Normal
REG	1	0.009	0.58	2.72	Normal
DEC	1	0.005	0.75	2.78	Normal
BJ	1	0.001	0.37	1.03	Normal
***srh0* (HW) = 3**	***srh0* (REG) = 2**	***srh0* (DEC) = 3**	***srh0* (BJ) = 3**		

Para todos os testes são exibidos *valor-p* (*pvalue*), limites de confiança inferior (*inf*) e superior *(sup)* para a mediana observada e status do teste de normalidade Jarque-Bera (*JB*) para a distribuição estatística das diferenças de desempenho: *normal* ou *não*.

Com base nos valores observados (Tabelas 48 e 49) para o indicador *srh0* (seção 4.4.2.2), pode-se chegar às conclusões da Tabela 50. Vale destacar também as melhores configurações observadas para o sistema NEW – previsores HW+ e HW- com referência SALY2, desempenhos aproximados MSE, geração de pesos históricos MQR, janela expansiva e normalização padrão (para a série DIESEL); previsores HW+, HW-, BJ+ e BJ- com referência SALY2, desempenhos aproximados MAD, geração de pesos históricos POOL, janela expansiva e normalização soma-1 (para a série GLP).

Tabela 50 – Conclusões para combinação NEW

	DIESEL		*GLP*	
Benchmarking	*srh0*	*Conclusão*	*srh0*	*Conclusão*
HW	3	A combinação é melhor	3	A combinação é melhor
REG	3	A combinação é melhor	2	A combinação é melhor
DEC	3	A combinação é melhor	3	A combinação é melhor
BJ	3	A combinação é melhor	3	A combinação é melhor
***srh0+* = 23**				

As Figuras 37 e 38 exibem a evolução dos pesos de combinação selecionados, ao longo do horizonte de teste. Em se tratando do sistema NEW, a Tabela 51 exibe informações das redes neurais responsáveis pelas ponderações.

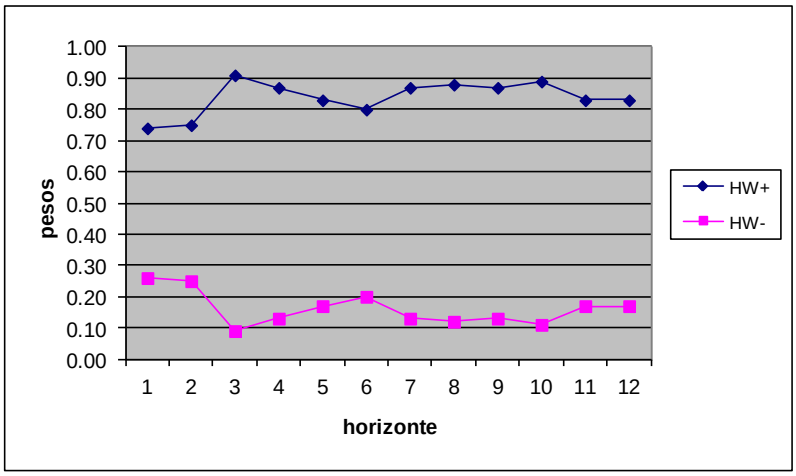

Figura 37 – Evolução dos pesos de combinação fora da amostra (DIESEL).

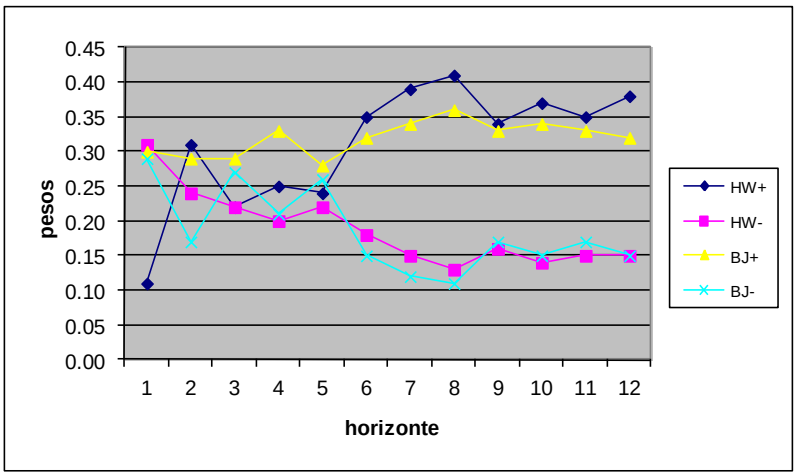

Figura 38 – Evolução dos pesos de combinação fora da amostra (GLP).

Tabela 51 – Redes neurais selecionadas

Arquitetura	*DIESEL*	*GLP*
Número de entradas	2	4
Número de saídas	2	4
Neurônios na camada escondida	23	19
Época de parada	61	146
Tempo de treinamento (h)	0.50	2.61
SMAPE estimação	3.57	3.81
SMAPE validação	3.82	2.85
SMAPE teste	2.07	2.32

Em todos os experimentos foram consideradas 9 replicações, com teste de 1 a 30 neurônios em cada uma (seção 3.9).

4.4.5. Análise comparativa

4.4.5.1. Comparação dirigida

As análises individuais conduzidas anteriormente tinham como objetivo comparar um método de combinação com seus previsores componentes. Na análise comparativa dirigida, comparam-se entre si os métodos de combinação testados – **tradicional**, **limiar** e **NEW** – de maneira restrita. Em termos práticos, a ideia é comparar a melhor combinação NEW (explorada individualmente na seção 4.4.4.2) com as melhores combinações dos tipos limiar e tradicional, quando estas últimas estiverem restritas ao mesmo escopo – previsores e métodos de geração de pesos – definido para o sistema NEW. O objetivo deste procedimento é estabelecer uma comparação sem **viés**, na medida em que o NEW, por decisão de projeto, foi testado apenas com dois previsores básicos (HW e BJ) e somente com métodos convexos para geração de pesos históricos (AFTER, BG1, MQR e AVG, este último através do método POOL) (seção 4.4.4.1).

As Tabelas 52 e 53 e as Figuras 39 e 40 exibem a evolução dos SMAPEs ao longo do horizonte de teste, considerando as combinações selecionadas para comparação dirigida (aquelas com menor erro composto, limitadas ao escopo dos experimentos NEW). Estas combinações são chamadas aqui de **combinações dirigidas**.

Tabela 52 – Desempenhos para comparação dirigida (DIESEL)

h	*TRADICIONAL DIRIGIDA*	*LIMIAR DIRIGIDA*	*NEW*
1	4.46	7.19	5.76
2	3.29	4.25	3.31
3	4.80	3.52	2.94
4	4.96	2.74	2.41
5	5.43	2.59	2.51
6	5.54	2.26	2.48
7	5.88	2.20	2.50
8	5.84	1.99	2.19
9	5.71	1.88	2.00
10	5.19	2.09	2.27
11	5.23	2.11	2.14
12	5.20	2.32	2.07

Melhor combinação tradicional dirigida: *2c, geração MQR, janela mínima & pesos dinâmicos;*

Melhor combinação limiar dirigida: *1a, geração MQR, janela mínima & pesos dinâmicos;*

Melhor combinação NEW: *new1a_mqr_mse, SALY2, janela expansiva & normalização padrão.*

Tabela 53 – Desempenhos para comparação dirigida (GLP)

h	*TRADICIONAL DIRIGIDA*	*LIMIAR DIRIGIDA*	*NEW*
1	2.14	3.98	0.91
2	2.06	3.38	1.88
3	3.27	3.49	3.20
4	3.66	3.22	3.38
5	3.31	2.70	3.11
6	3.53	2.53	3.04
7	3.51	2.34	2.70
8	3.62	2.44	2.51
9	3.49	2.39	2.28
10	3.21	2.16	2.36
11	3.44	2.67	2.46
12	3.40	2.93	2.32

Melhor combinação tradicional dirigida: *2c, geração AFTER, janela mínima & pesos dinâmicos;*

Melhor combinação limiar dirigida: *22c, geração BG1, janela mínima & pesos dinâmicos;*

Melhor combinação NEW: *new2c_pool_mad, SALY2, janela expansiva & normalização soma-1.*

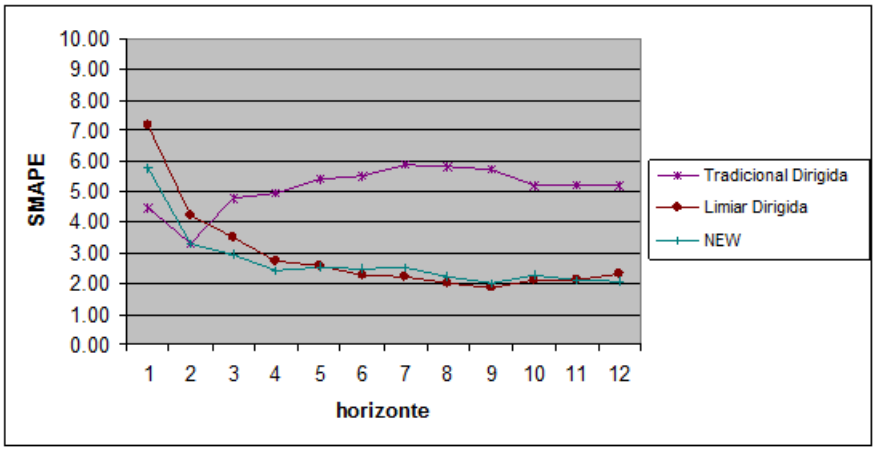

Figura 39 – Evolução dos SMAPEs fora da amostra (DIESEL).

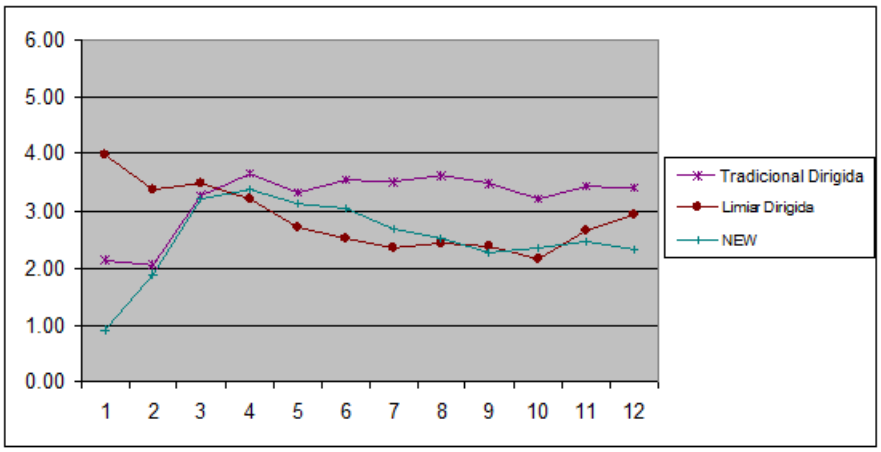

Figura 40 – Evolução dos SMAPEs fora da amostra (GLP).

A Tabela 54 atualiza as conclusões da Tabela 23, obtidas para as combinações tradicionais **livres**[15], e as Figuras 41 e 42 exibem a evolução dos pesos (fora da amostra) para as melhores combinações tradicionais **dirigidas** (Tabelas 52 e 53). Na Figura 42 só há variação efetiva dos pesos entre os dois primeiros passos do horizonte; isto se explica por uma característica particular do método AFTER (seção 2.4.6): pesos zerados permanecem assim ao longo de todo o horizonte subsequente. Como neste caso as melhores combinações limiares dirigidas (Tabelas 52 e 53) foram iguais às melhores combinações limiares livres (Tabelas 29 e 30), as conclusões da Tabela 35 permanecem inalteradas e as respectivas evoluções de pesos também podem ser omitidas, por já terem sido

[15] O termo **combinação livre** é definido na seção 4.4.5.2.

exibidas na seção 4.4.3.2. Também, as evoluções de pesos para as melhores combinações NEW podem ser vistas na seção 4.4.4.2.

Tabela 54 – Conclusões para combinação tradicional dirigida

	DIESEL		*GLP*	
Benchmarking	*srh0*	*Conclusão*	*srh0*	*Conclusão*
HW	2	Indiferente	3	A combinação é melhor
REG	3	A combinação é melhor	2	A combinação é melhor
DEC	3	A combinação é melhor	2	A combinação é melhor
BJ	0	Indiferente	-3	A combinação é pior
***srh0*+ = 15**				

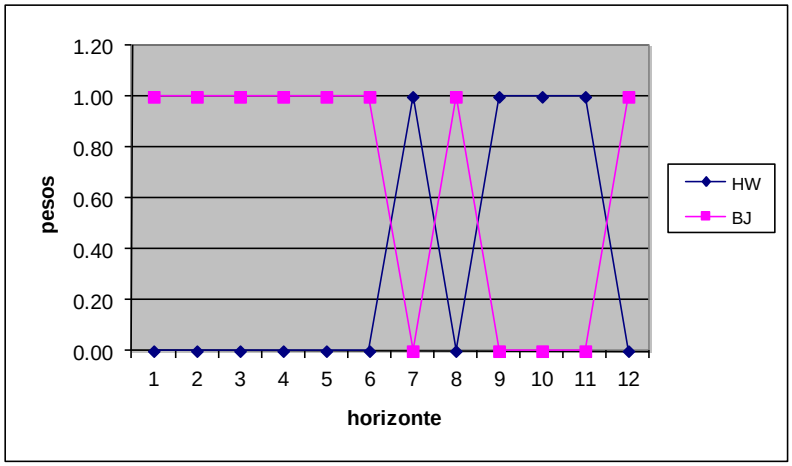

Figura 41 – Evolução dos pesos de combinação fora da amostra (DIESEL).

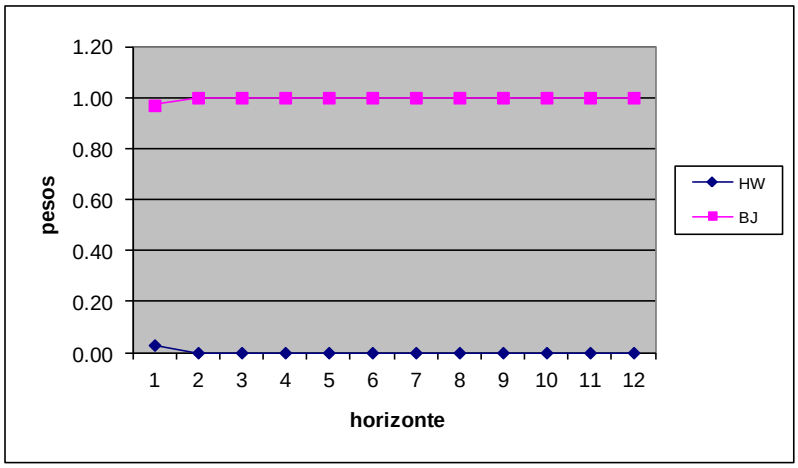

Figura 42 – Evolução dos pesos de combinação fora da amostra (GLP).

As Tabelas 55 e 56 exibem as diferenças de desempenho tomadas período a período, fora da amostra, na ordem "erro do previsor combinado (*benchmarking*) **menos** erro do previsor NEW". Como em seções anteriores, todos os testes de hipótese sugeridos para comparação de desempenhos foram executados; as Tabelas 57 e 58 exibem os resultados.

Tabela 55 – Diferenças de desempenho combinações/NEW (DIESEL)

h	*TRADICIONAL DIRIGIDA*	*LIMIAR DIRIGIDA*
1	-1.30	1.43
2	-0.02	0.95
3	1.86	0.59
4	2.55	0.32
5	2.92	0.08
6	3.06	-0.22
7	3.38	-0.30
8	3.65	-0.20
9	3.71	-0.12
10	2.92	-0.18
11	3.09	-0.03
12	3.13	0.25
MEDIANA	**2.99**	**0.03**

Tabela 56 – Diferenças de desempenho combinações/NEW (GLP)

h	*TRADICIONAL DIRIGIDA*	*LIMIAR DIRIGIDA*
1	1.23	3.07
2	0.18	1.50
3	0.08	0.30
4	0.28	-0.16
5	0.20	-0.42
6	0.49	-0.51
7	0.82	-0.35
8	1.11	-0.07
9	1.22	0.11
10	0.85	-0.20
11	0.98	0.21
12	1.08	0.61
MEDIANA	**0.83**	**0.02**

Tabela 57 – Testes de hipótese (DIESEL)

NEW →	*new1a_mqr_mse, SALY2, janela expansiva & normalização padrão*				
Teste t					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
TRAD. DIRIGIDA	1	0.000	1.43	3.39	Não normal
LIMIAR DIRIGIDA	0	0.193	-0.13	0.55	Normal
Teste de sinais					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
TRAD. DIRIGIDA	1	0.039	1.86	3.38	Não normal
LIMIAR DIRIGIDA	0	1.000	-0.20	0.58	Normal
Teste de Wilcoxon					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
TRAD. DIRIGIDA	1	0.002	1.21	3.32	Não normal
LIMIAR DIRIGIDA	0	0.380	-0.16	0.61	Normal
***srh0* (TRAD. DIRIGIDA) = 3**		***srh0* (LIMIAR DIRIGIDA) = 0**			

Para todos os testes são exibidos *valor-p* (*pvalue*), limites de confiança inferior (*inf*) e superior *(sup)* para a mediana observada e status do teste de normalidade Jarque-Bera (*JB*) para a distribuição estatística das diferenças de desempenho: *normal* ou *não*.

Tabela 58 – Testes de hipótese (GLP)

NEW →	*new2c_pool_mad, SALY2, janela expansiva & normalização soma 1*				
Teste t					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
TRAD. DIRIGIDA	1	0.000	0.43	0.99	Normal
LIMIAR DIRIGIDA	0	0.272	-0.31	0.99	Não normal
Teste de sinais					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
TRAD. DIRIGIDA	1	0.001	0.20	1.11	Normal
LIMIAR DIRIGIDA	0	1.000	-0.36	0.61	Não normal
Teste de Wilcoxon					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
TRAD. DIRIGIDA	1	0.001	0.39	1.04	Normal
LIMIAR DIRIGIDA	0	0.622	-0.24	1.06	Não normal
***srh0* (TRAD. DIRIGIDA) = 3**		***srh0* (LIMIAR DIRIGIDA) = 0**			

Para todos os testes são exibidos *valor-p* (*pvalue*), limites de confiança inferior (*inf*) e superior *(sup)* para a mediana observada e status do teste de normalidade Jarque-Bera (*JB*) para a distribuição estatística das diferenças de desempenho: *normal* ou *não*.

Com base nos valores observados (Tabelas 57 e 58) para o indicador *srh0* (seção 4.4.2.2), pode-se chegar às conclusões da Tabela 59. Embora as comparações entre as combinações NEW e limiar estejam sinalizadas em ambas as séries como "indiferente", a inspeção das Tabelas 35 e 50 mostra que o NEW leva vantagem, na medida em que apresenta melhores indicadores quando comparado aos previsores individuais: o *srh0+* vale **16** para a combinação limiar (Tabela 35) e **23** para o NEW (Tabela 50) (o máximo valor possível para o *srh0+* seria **24**).

Tabela 59 – Conclusões para combinação NEW

	DIESEL		*GLP*	
Benchmarking	*srh0*	*Conclusão*	*srh0*	*Conclusão*
TRAD.DIRIGIDA	3	O NEW é melhor	3	O NEW é melhor
LIMIAR DIRIGIDA	0	Indiferente	0	Indiferente

Por fim, a Tabela 60 exibe diferentes métricas de desempenho (calculadas fora da amostra) para os experimentos selecionados na comparação dirigida.

Tabela 60 – Métricas de desempenho total na comparação dirigida

	DIESEL			*GLP*		
Experimento	*SMAPE*	*RAE*	*UTHEIL*	*SMAPE*	*RAE*	*UTHEIL*
TRAD. DIRIGIDA	5.20	0.27	0.47	3.40	0.67	0.52
LIMIAR DIRIGIDA	2.32	0.12	0.24	2.93	0.57	0.49
NEW	2.07	0.11	0.22	2.32	0.45	0.38

4.4.5.2. Comparação livre

Aqui, os melhores resultados para cada método de combinação (explorados individualmente nas seções 4.4.2.2, 4.4.3.2 e 4.4.4.2) são comparados livremente entre si, sem qualquer restrição. As Tabelas 61 e 62 e as Figuras 43 e 44 exibem a evolução dos SMAPEs ao longo do horizonte de teste, considerando as combinações selecionadas para comparação livre (aquelas com menor erro composto). Estas combinações são chamadas aqui de **combinações livres**.

Tabela 61 – Desempenhos para comparação livre (DIESEL)

h	*TRADICIONAL*	*LIMIAR*	*NEW*
1	8.20	7.19	5.76
2	4.92	4.25	3.31
3	4.56	3.52	2.94
4	3.80	2.74	2.41
5	3.72	2.59	2.51
6	3.46	2.26	2.48
7	3.51	2.20	2.50
8	3.29	1.99	2.19
9	3.01	1.88	2.00
10	3.00	2.09	2.27
11	2.82	2.11	2.14
12	2.65	2.32	2.07

Melhor combinação tradicional: *2e, geração MQI, janela mínima & pesos estáticos;*

Melhor combinação limiar: *1a, geração MQR, janela mínima & pesos dinâmicos;*

Melhor combinação NEW: *new1a_mqr_mse, SALY2, janela expansiva & normalização padrão.*

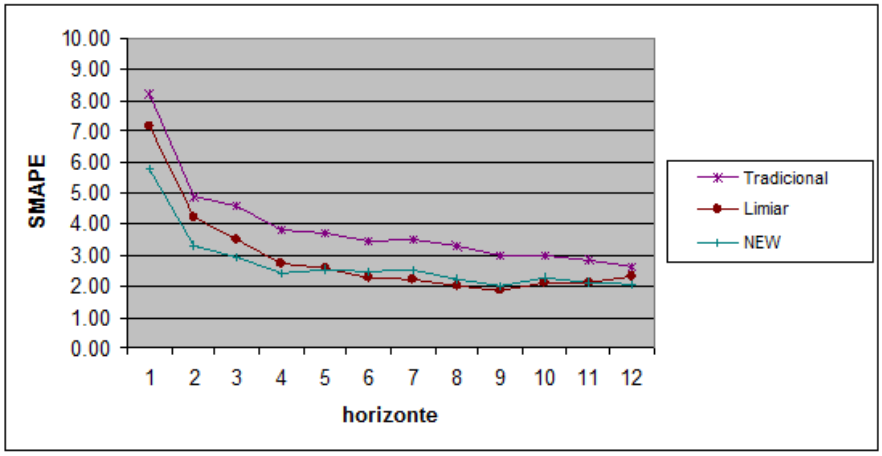

Figura 43 – Evolução dos SMAPEs fora da amostra (DIESEL).

Tabela 62 – Desempenhos para comparação livre (GLP)

h	*TRADICIONAL*	*LIMIAR*	*NEW*
1	1.42	3.98	0.91
2	1.61	3.38	1.88
3	3.11	3.49	3.20
4	3.41	3.22	3.38
5	3.15	2.70	3.11
6	3.39	2.53	3.04
7	3.41	2.34	2.70
8	3.58	2.44	2.51
9	3.46	2.39	2.28
10	3.22	2.16	2.36
11	3.45	2.67	2.46
12	3.36	2.93	2.32

Melhor combinação tradicional: *2e, geração SMQI, janela mínima & pesos estáticos;*

Melhor combinação limiar: *22c, geração BG1, janela mínima & pesos dinâmicos;*

Melhor combinação NEW: *new2c_pool_mad, SALY2, janela expansiva & normalização soma 1.*

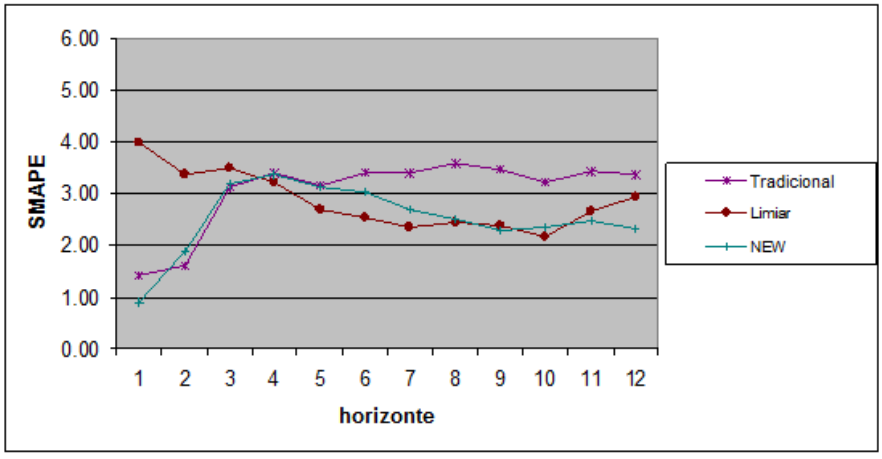

Figura 44 – Evolução dos SMAPEs fora da amostra (GLP).

As Tabelas 63 e 64 exibem as diferenças de desempenho tomadas período a período, fora da amostra, na ordem "erro do previsor combinado (*benchmarking*) **menos** erro do previsor NEW". Como em seções anteriores, todos os testes de hipótese sugeridos para comparação de desempenhos foram executados; as Tabelas 65 e 66 exibem os resultados.

Tabela 63 – Diferenças de desempenho combinações/NEW (DIESEL)

h	*TRADICIONAL*	*LIMIAR*
1	2.44	1.43
2	1.62	0.95
3	1.63	0.59
4	1.39	0.32
5	1.21	0.08
6	0.98	-0.22
7	1.01	-0.30
8	1.09	-0.20
9	1.01	-0.12
10	0.73	-0.18
11	0.69	-0.03
12	0.58	0.25
MEDIANA	**1.05**	**0.03**

Tabela 64 – Diferenças de desempenho combinações/NEW (GLP)

h	*TRADICIONAL*	*LIMIAR*
1	0.51	3.07
2	-0.27	1.50
3	-0.09	0.30
4	0.04	-0.16
5	0.04	-0.42
6	0.36	-0.51
7	0.72	-0.35
8	1.07	-0.07
9	1.18	0.11
10	0.86	-0.20
11	0.98	0.21
12	1.04	0.61
MEDIANA	**0.61**	**0.02**

Tabela 65 – Testes de hipótese (DIESEL)

NEW →	*new1a_mqr_mse, SALY2, janela expansiva & normalização padrão*				
Teste t					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
TRADICIONAL	1	0.000	0.87	1.52	Normal
LIMIAR	0	0.193	-0.13	0.55	Normal
Teste de sinais					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
TRADICIONAL	1	0.001	0.73	1.61	Normal
LIMIAR	0	1.000	-0.20	0.58	Normal
Teste de Wilcoxon					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
TRADICIONAL	1	0.001	0.86	1.56	Normal
LIMIAR	0	0.380	-0.16	0.61	Normal
***srh0* (TRADICIONAL) = 3**		***srh0* (LIMIAR) = 0**			

Para todos os testes são exibidos *valor-p* (*pvalue*), limites de confiança inferior (*inf*) e superior *(sup)* para a mediana observada e status do teste de normalidade Jarque-Bera (*JB*) para a distribuição estatística das diferenças de desempenho: *normal* ou *não*.

Tabela 66 – Testes de hipótese (GLP)

NEW →	*new2c_pool_mad, SALY2, janela expansiva & normalização soma 1*				
Teste t					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
TRADICIONAL	1	0.000	0.21	0.86	Normal
LIMIAR	0	0.270	-0.31	0.99	Não normal
Teste de sinais					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
TRADICIONAL	1	0.040	0.03	1.04	Normal
LIMIAR	0	1.000	-0.36	0.61	Não normal
Teste de Wilcoxon					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
TRADICIONAL	1	0.010	0.19	0.95	Normal
LIMIAR	0	0.620	-0.24	1.06	Não normal
***srh0* (TRADICIONAL) = 3**		***srh0* (LIMIAR) = 0**			

Para todos os testes são exibidos *valor-p* (*pvalue*), limites de confiança inferior (*inf*) e superior *(sup)* para a mediana observada e status do teste de normalidade Jarque-Bera (*JB*) para a distribuição estatística das diferenças de desempenho: *normal* ou *não*.

Com base nos valores observados (Tabelas 65 e 66) para o indicador *srh0* (seção 4.4.2.2), pode-se chegar às conclusões da Tabela 67. A exemplo da seção 4.4.5.1, embora as comparações entre as combinações NEW e limiar estejam sinalizadas em ambas as séries como "indiferente", a inspeção das Tabelas 35 e 50 mostra que o NEW leva vantagem quando comparado aos previsores individuais: o *srh0+* vale **16** para a combinação limiar[16] e **23** para o NEW.

Tabela 67 – Conclusões para combinação NEW

	DIESEL		*GLP*	
Benchmarking	*srh0*	*Conclusão*	*srh0*	*Conclusão*
TRADICIONAL	3	O NEW é melhor	3	O NEW é melhor
LIMIAR	0	Indiferente	0	Indiferente

Por fim, a Tabela 68 exibe diferentes métricas de desempenho (calculadas fora da amostra) para os experimentos selecionados na comparação livre.

Tabela 68 – Métricas de desempenho total na comparação livre

	DIESEL			*GLP*		
Experimento	*SMAPE*	*RAE*	*UTHEIL*	*SMAPE*	*RAE*	*UTHEIL*
TRADICIONAL	2.65	0.14	0.26	3.36	0.67	0.52
LIMIAR	2.32	0.12	0.24	2.93	0.57	0.49
NEW	2.07	0.11	0.22	2.32	0.45	0.38

4.4.5.3. Tempo de processamento

A Tabela 69 exibe as configurações de hardware e software utilizadas neste trabalho e os tempos unitários de processamento para os principais experimentos conduzidos, quando aplicados em uma única série temporal (e.g. DIESEL).

[16] Para este estudo, as melhores combinações limiares **livres** foram iguais às melhores combinações limiares **dirigidas**.

Tabela 69 – Tempos unitários de processamento (s)

Configuração	*Limiar1*	*Limiar2*	*NEW1*	*NEW2*
Intel Core 2 Quad 2.66Ghz 8GB RAM Windows 7 Home Premium MATLAB R2008b	2	3	1858	4659
Intel Core 2 Duo 3.49GB RAM Windows XP Professional MATLAB R14	2	3	2217	8086
MÉDIA	**2**	**3**	**2038**	**6372**

Tempos unitários (em segundos) para experimentos de combinação limiar com 2 ou 4 previsores (*Limiar1* e *Limiar2*) e respectivos experimentos NEW (*NEW1 e NEW2*).

Pelos tempos médios na Tabela 69 (em segundos) é possível estimar o tempo total de processamento para os experimentos realizados, admitindo-se execução sequencial. No caso das combinações limiares, foram realizados 80 experimentos por série, sendo 64 do tipo Limiar1 (1a, 1b, 1c e 1d) e 16 do tipo Limiar2 (22c) (seção 4.4.3.1); desta forma, para as séries DIESEL e GLP juntas, tem-se um tempo total estimado de (aproximadamente) **6 minutos**. Já no caso das combinações NEW, foram executados 256 experimentos por série, sendo 128 do tipo NEW1 (prefixo "new1a") e 128 do tipo NEW2 (prefixo "new2c") (seção 4.4.4.1); para as séries DIESEL e GLP juntas, o tempo total estimado é de (aproximadamente) **25 dias**.

Em todos os experimentos de redes neurais neste trabalho foram consideradas 9 replicações, com teste de 1 a 30 neurônios em cada uma (seção 3.9). Isto significa um total de **270** configurações de rede por experimento – sendo apenas uma selecionada ao final – o que deixa as comparações de tempo entre as combinações limiares e os modelos NEW um tanto **enviesadas**: em experimentos futuros, o número de replicações pode até ser mantido, mas a busca pelo melhor número de neurônios pode ser limitada a um intervalo menor. Em um cenário otimista, onde se admita o uso de apenas uma arquitetura por experimento, o tempo de treinamento das redes neurais cairia 270 vezes, e os 25 dias de processamento seriam reduzidos a (aproximadamente) **2hs**.

4.4.5.4. Resumo

Os resultados da comparação **dirigida** (seção 4.4.5.1) trazem indícios razoáveis de que o sistema NEW pode agregar valor aos procedimentos tradicionais. Esta conclusão vem da análise conjunta dos resultados da Tabela 59 e do indicador *srh0+* (seção 4.4.2.2), que apresenta valores relativamente altos para cada um dos métodos de combinação: **15** para a combinação tradicional, **16** para a combinação limiar e **23** para a combinação NEW, todos numa escala que vai até **24** (Tabelas 54, 35 e 50). De fato, estes resultados reforçam uma conclusão recorrente na literatura: há vantagem prática em combinar previsores (seção 1.3). Mesmo na comparação **livre** (seção 4.4.5.2), pelo mesmo tipo de análise (agora com as Tabelas 67, 23, 35 e 50), este indicativo permanece. Observando-se isoladamente os desempenhos totais fora da amostra (acumulados em 12 meses), o sistema NEW também apresentou os melhores resultados: SMAPEs de **2.07** para DIESEL e **2.32** para GLP (Tabela 68).

Como desvantagem do sistema proposto neste trabalho, pode-se citar o elevado tempo de **treinamento**: os modelos do tipo NEW1 têm tempos unitários de treinamento que variam entre **4** (cenário otimista) e **1000** (cenário conservador) vezes o tempo unitário do método tradicional correspondente (Limiar1); já os modelos do tipo NEW2 multiplicam o tempo unitário do método tradicional (Limiar2) por um valor entre **8** (cenário otimista) e **2000** (cenário conservador) (Tabela 69). Apesar disso, deve-se considerar o fato de que, na grande maioria dos casos, o treinamento é uma atividade de baixa frequência, sendo realizada apenas quando uma grande massa de dados novos (séries realizadas) puder ser reunida (e.g. a cada seis meses ou mesmo anualmente).

4.5. CASO 2: Competição NN3

A competição NN3 foi uma competição entre métodos de previsão conduzida durante os anos de 2006 e 2007, criada essencialmente para avaliar algoritmos baseados em redes neurais ou inteligência computacional (NN3, 2011). Competições deste tipo têm sido continuamente realizadas na comunidade (IJCNN, 2011; ICTSF, 2012).

O banco de dados da NN3 é constituído, na sua forma completa, de **111** séries temporais mensais, relacionadas a atividades de transporte tais como: tráfego em autoestradas, tráfego de carros em túneis, tráfego em pedágios, tráfego de pessoas em estações de metrô, voos domésticos, entregas de importação, cruzamento de fronteiras, fluxo em dutos e transporte ferroviário. Há também uma versão **reduzida** do banco, que reúne apenas as últimas **11** séries do conjunto (listadas no Apêndice C). Esta foi a versão utilizada neste trabalho.

Basicamente, três passos deviam ser seguidos para participar da NN3:

1. Desenvolver um método de previsão bem documentado, que possa ser automatizado;
2. Testar o método em todas as séries do banco de dados fornecido pelos organizadores, produzindo, para cada série de tamanho τ, previsões para $\tau+1$, $\tau+2$, ..., $\tau+18$;
3. Submeter as previsões geradas para julgamento.

Conhecedores das realizações fora da amostra, os julgadores puderam calcular, para cada competidor, a média (para todas as séries) dos desempenhos SMAPE (73) acumulados **18** passos a frente. Foi declarado vencedor o método (de inteligência computacional) com melhor desempenho médio.

Deve-se observar que nas competições tradicionais não há testes estatísticos[17] para comparação dos resultados obtidos pelos competidores, sendo este, inclusive, um motivo de crítica (Stekler, 2001). Por outro lado, como será visto mais adiante, a disputa pelo posto de **campeão** é acirrada, girando em torno da segunda casa decimal; parece então bastante provável que os primeiros métodos sejam estatisticamente equivalentes, o que deixaria o ranking com muitos primeiros colocados, esvaziando a competição em sua característica de estimular a participação de estudantes e pesquisadores.

[17] Neste trabalho **não** são realizados testes estatísticos para comparar vencedores da competição NN3, mas sim para comparar os principais métodos de combinação testados entre si e entre seus previsores componentes.

4.5.1.
Previsores individuais

Neste estudo de caso, apenas as três primeiras metodologias de previsão sugeridas na seção 4.2 foram utilizadas (HW, REG e DEC). A metodologia Box & Jenkins foi descartada por dois motivos: (i) dar característica mais automática às técnicas de combinação testadas, dispensando o ajuste pormenorizado exigido pela metodologia BJ às 11 séries da competição; (ii) reduzir convenientemente o tempo de processamento. Para cada série, em concordância com o regulamento da competição, separou-se sempre os últimos **18** meses de dados para teste.

Com os modelos ajustados, foram geradas previsões até 18 passos a frente, de maneira **não recursiva**, i.e., sem reestimação de parâmetros a cada passo. A Tabela 70 exibe os desempenhos médios **totais** – considerando sempre as 11 séries da competição – obtidos dentro e fora da amostra (18 meses); o previsor DEC apresentou o menor **erro médio composto** (SMAPE médio amostra + SMAPE médio teste).

Tabela 70 – Desempenhos médios totais

Método	*SMAPE Amostra*	*SMAPE Teste*
HW	19.49	15.08
REG	15.91	15.74
DEC	14.10	16.25

SMAPEs médios para as 11 séries da competição.

A Tabela 71 exibe os desempenhos totais **por série**, medidos fora da amostra; as Figuras 45 e 46 ilustram estes desempenhos com gráficos de barras e diagramas de dispersão (*boxplots*).

Tabela 71 – Desempenhos totais por série (fora da amostra)

série	*HW*	*REG*	*DEC*
1	1.80	2.67	2.61
2	30.63	24.55	29.01
3	29.60	33.81	29.16
4	6.11	9.37	13.78
5	1.53	10.22	10.66
6	5.38	3.62	4.10
7	5.10	3.48	3.29
8	36.29	29.54	29.62
9	7.77	16.88	16.69
10	30.48	27.25	28.62
11	11.14	11.79	11.23

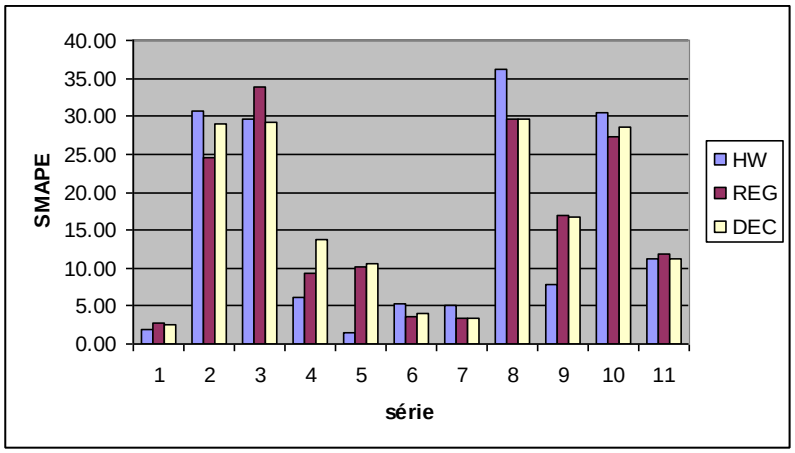

Figura 45 – SMAPEs fora da amostra: gráficos de barras.

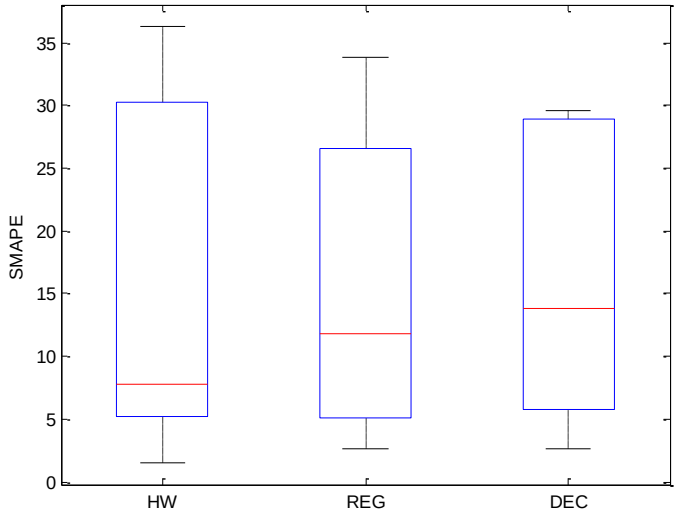

Figura 46 – SMAPEs fora da amostra: *boxplots*.

Embora com mediana mais elevada ao longo das 11 séries, o previsor DEC teve a menor dispersão de erro (Figura 46). A Tabela 72 e a Figura 47 exibem a evolução dos SMAPEs médios no conjunto de teste, ao longo do horizonte (*h*) de 18 meses.

Tabela 72 – Evolução dos SMAPEs médios fora da amostra

h	*HW*	*REG*	*DEC*
1	10.66	13.45	13.73
2	9.32	13.04	13.80
3	8.59	13.00	13.05
4	8.71	12.84	12.59
5	9.39	13.48	13.00
6	10.07	13.40	13.58
7	10.78	14.30	14.35
8	11.03	14.76	14.34
9	11.09	14.89	14.40
10	11.14	14.80	14.21
11	12.17	15.24	14.40
12	12.98	15.63	14.88
13	13.88	16.47	15.64
14	14.19	16.58	15.80
15	14.32	16.12	15.78
16	14.73	15.90	16.02
17	14.82	15.67	16.04
18	15.08	15.74	16.25

SMAPEs médios para as 11 séries da competição.

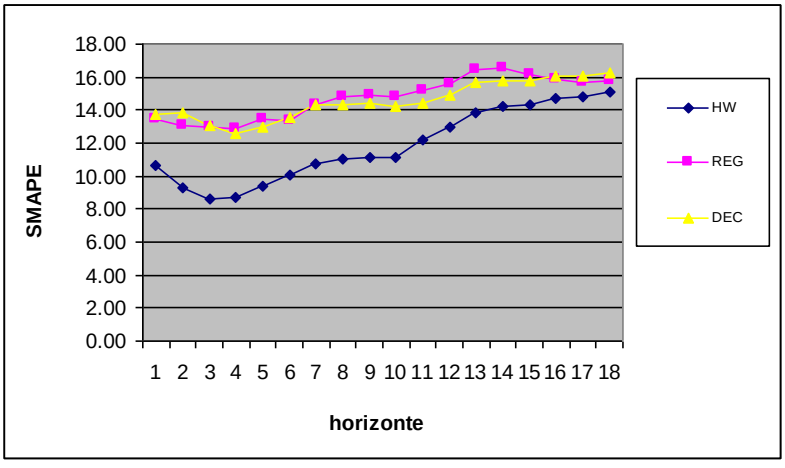

Figura 47 – Evolução dos SMAPEs médios fora da amostra (NN3).

4.5.2.
Combinações tradicionais

4.5.2.1.
Experimentos

A exemplo da seção 4.4.2.1, foi definida a sequência de experimentos da Tabela 73, organizada de acordo com os previsores sendo combinados; como se observa, todas as possíveis combinações foram testadas.

Tabela 73 – Experimentos de combinação tradicional

Experimento	*Previsores Combinados*
2a	HW, REG
2b	HW, DEC
2d	REG, DEC
3a	HW, REG, DEC

Dadas as 4 combinações da Tabela 73 e considerando-se as mesmas variações de hiperparâmetros da seção 4.4.2.1 – 11 métodos para geração de pesos, 2 tamanhos da janela (mínima e expansiva) e 2 tipos de geração (dinâmica ou estática) – o total de experimentos foi de **176** por série. As Tabelas 74 a 77 exibem os resultados (médios) selecionados, obtidos nos conjuntos de **validação** (constituídos, para cada série, pelos últimos 18 meses da amostra, imediatamente anteriores ao conjunto de teste) e **teste** (18 meses fora da amostra). Em todos os casos, o método de geração de pesos foi escolhido pelo melhor desempenho (médio) na validação; os valores em negrito indicam os experimentos com menor erro médio composto (SMAPE médio validação + SMAPE médio teste).

Tabela 74 – Desempenhos médios totais para *janela mínima & pesos dinâmicos*

Experimento	*Método*	*SMAPE Validação*	*SMAPE Teste*
2a	SAFTER	12.07	14.09
2b	AFTER	11.14	15.41
2d	SAFTER	11.16	15.02
3a	AFTER	11.05	14.65

SMAPEs médios para as 11 séries da competição. Para cada experimento, o método de geração de pesos selecionado foi o de melhor desempenho (médio) na validação.

Tabela 75 – Desempenhos médios totais para *janela mínima & pesos estáticos*

Experimento	*Método*	*SMAPE Validação*	*SMAPE Teste*
2a	SAFTER	12.09	13.92
2b	AFTER	11.15	14.68
2d	SAFTER	11.10	15.02
3a	AFTER	10.49	14.15

SMAPEs médios para as 11 séries da competição. Para cada experimento, o método de geração de pesos selecionado foi o de melhor desempenho (médio) na validação.

Tabela 76 – Desempenhos médios totais para *janela expansiva & pesos dinâmicos*

Experimento	*Método*	*SMAPE Validação*	*SMAPE Teste*
2a	SBG2	11.98	14.04
2b	BG2	10.49	15.29
2d	SMQR	11.18	14.98
3a	BG2	10.86	15.13

SMAPEs médios para as 11 séries da competição. Para cada experimento, o método de geração de pesos selecionado foi o de melhor desempenho (médio) na validação.

Tabela 77 – Desempenhos médios totais para *janela expansiva & pesos estáticos*

Experimento	*Método*	*SMAPE Validação*	*SMAPE Teste*
2a	SAFTER	12.01	13.63
2b	BG2	10.01	14.38
2d	SAFTER	11.05	15.00
3a	AFTER	**10.08**	**13.71**

SMAPEs médios para as 11 séries da competição. Para cada experimento, o método de geração de pesos selecionado foi o de melhor desempenho (médio) na validação.

4.5.2.2. Análise individual

A Tabela 78 e a Figura 48 exibem a evolução dos SMAPEs médios ao longo do horizonte de teste, considerando os previsores individuais e a melhor combinação tradicional obtida (de menor erro médio composto).

Tabela 78 – Evolução dos SMAPEs médios fora da amostra (NN3)

h	*HW*	*REG*	*DEC*	*COMBINAÇÃO*
1	10.66	13.45	13.73	9.47
2	9.32	13.04	13.80	9.96
3	8.59	13.00	13.05	9.39
4	8.71	12.84	12.59	8.87
5	9.39	13.48	13.00	9.18
6	10.07	13.40	13.58	9.70
7	10.78	14.30	14.35	10.44
8	11.03	14.76	14.34	10.65
9	11.09	14.89	14.40	10.79
10	11.14	14.80	14.21	10.80
11	12.17	15.24	14.40	11.50
12	12.98	15.63	14.88	12.12
13	13.88	16.47	15.64	13.16
14	14.19	16.58	15.80	13.48
15	14.32	16.12	15.78	13.41
16	14.73	15.90	16.02	13.63
17	14.82	15.67	16.04	13.55
18	15.08	15.74	16.25	13.71

Melhor combinação tradicional: *3a, geração AFTER, janela expansiva & pesos estáticos.*
SMAPEs médios para as 11 séries da competição.

A Tabela 79 exibe as diferenças de desempenho médio tomadas período a período, fora da amostra, na ordem "erro médio do previsor individual (*benchmarking*) **menos** erro médio do previsor combinado"; quanto mais positiva a diferença, melhor o método à direita da comparação.

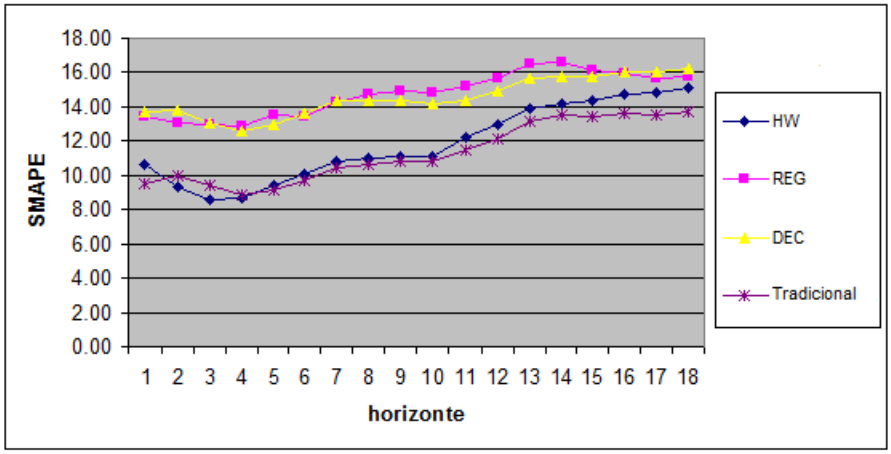

Figura 48 – Evolução dos SMAPEs médios fora da amostra (NN3).

Tabela 79 – Diferenças de desempenho médio individuais/combinação (NN3)

h	*HW*	*REG*	*DEC*
1	1.19	3.98	4.26
2	-0.64	3.08	3.84
3	-0.80	3.61	3.66
4	-0.16	3.97	3.72
5	0.21	4.30	3.82
6	0.37	3.70	3.88
7	0.34	3.86	3.91
8	0.38	4.11	3.69
9	0.30	4.10	3.61
10	0.34	4.00	3.41
11	0.67	3.74	2.90
12	0.86	3.51	2.76
13	0.72	3.31	2.48
14	0.71	3.10	2.32
15	0.91	2.71	2.37
16	1.10	2.27	2.39
17	1.27	2.12	2.49
18	1.37	2.03	2.54
MEDIANA	**0.53**	**3.66**	**3.51**

Calculadas as diferenças, todos os testes de hipótese sugeridos na seção 4.3.2 foram executados, sempre com o objetivo de verificar se as medianas das mesmas são diferentes de zero (i.e., se há diferença significativa entre os métodos comparados). A Tabela 80 exibe os resultados.

Tabela 80 – Testes de hipótese (NN3)

Combinação →	*3a, geração AFTER, janela expansiva & pesos estáticos.*				
Teste t					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	1	0.000	0.21	0.81	Normal
REG	1	0.000	3.06	3.77	Normal
DEC	1	0.000	2.89	3.56	Normal
Teste de sinais					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	1	0.010	0.30	0.91	Normal
REG	1	0.000	3.08	3.98	Normal
DEC	1	0.000	2.49	3.82	Normal
Teste de Wilcoxon					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	1	0.010	0.26	0.82	Normal
REG	1	0.000	3.04	3.86	Normal
DEC	1	0.000	2.95	3.69	Normal
***srh0* (HW) = 3** ***srh0* (REG) = 3** ***srh0* (DEC) = 3**					

Para todos os testes são exibidos *valor-p* (*pvalue*), limites de confiança inferior (*inf*) e superior *(sup)* para a mediana observada e status do teste de normalidade Jarque-Bera (*JB*) para a distribuição estatística das diferenças de desempenho: *normal* ou *não*.

Com base nos valores observados (Tabela 80) para o indicador *srh0* (seção 4.4.2.2), pode-se chegar às conclusões da Tabela 81.

Tabela 81 – Conclusões para combinação tradicional (NN3)

Benchmarking	*srh0*	*Conclusão*
HW	3	A combinação é melhor
REG	3	A combinação é melhor
DEC	3	A combinação é melhor
***srh0*+ = 9**		

4.5.3.
Combinações limiares

4.5.3.1.
Experimentos

A exemplo da seção 4.4.3.1, foi definida a sequência de experimentos da Tabela 82, nos mesmos moldes da Tabela 73. Além das combinações com previsores limiares derivados de um único modelo, testou-se uma combinação (22a) com 4 previsores limiares, derivados dos dois modelos de melhor desempenho no teste (HW e REG, Tabela 70).

Tabela 82 – Experimentos de combinação limiar

Experimento	*Previsores Combinados*
1a	HW+, HW-
1b	REG+, REG-
1c	DEC+, DEC-
22a	HW+, HW-, REG+, REG-

Dadas as 4 combinações da Tabela 82 e considerando-se as mesmas variações de hiperparâmetros da seção 4.4.3.1 – 4 métodos para geração de pesos convexos, 2 tamanhos da janela (mínima e expansiva) e 2 tipos de geração (dinâmica ou estática) – o total de experimentos foi de **64** por série. As Tabelas 83 a 86 exibem os resultados (médios) selecionados, obtidos dentro e fora da amostra (validação e teste). Como na seção 4.5.2.1, o método de geração de pesos foi escolhido pelo melhor desempenho (médio) na validação. Os valores em negrito indicam os experimentos com menor erro médio composto (SMAPE médio validação + SMAPE médio teste).

Tabela 83 – Desempenhos médios totais para *janela mínima & pesos dinâmicos*

Experimento	*Método*	*SMAPE Validação*	*SMAPE Teste*
1a	MQR	30.67	25.10
1b	BG1	17.23	20.47
1c	MQR	17.46	17.69
22a	MQR	19.56	17.50

SMAPEs médios para as 11 séries da competição. Para cada experimento, o método de geração de pesos selecionado foi o de melhor desempenho (médio) na validação.

Tabela 84 – Desempenhos médios totais para *janela mínima & pesos estáticos*

Experimento	*Método*	*SMAPE Validação*	*SMAPE Teste*
1a	MQR	27.79	18.65
1b	AVG	13.29	15.74
1c	AVG	**11.39**	**16.25**
22a	AVG	14.00	13.98

SMAPEs médios para as 11 séries da competição. Para cada experimento, o método de geração de pesos selecionado foi o de melhor desempenho (médio) na validação.

Tabela 85 – Desempenhos médios totais para *janela expansiva & pesos dinâmicos*

Experimento	*Método*	*SMAPE Validação*	*SMAPE Teste*
1a	MQR	29.61	20.77
1b	MQR	15.04	16.05
1c	MQR	12.50	16.19
22a	BG1	13.92	15.01

SMAPEs médios para as 11 séries da competição. Para cada experimento, o método de geração de pesos selecionado foi o de melhor desempenho (médio) na validação.

Tabela 86 – Desempenhos médios totais para *janela expansiva & pesos estáticos*

Experimento	*Método*	*SMAPE Validação*	*SMAPE Teste*
1a	MQR	26.02	20.78
1b	AVG	13.29	15.74
1c	AVG	**11.39**	**16.25**
22a	BG1	13.37	14.78

SMAPEs médios para as 11 séries da competição. Para cada experimento, o método de geração de pesos selecionado foi o de melhor desempenho (médio) na validação.

Nos experimentos acima, verifica-se que o método da média entre DEC+ e DEC- produz o melhor resultado, o que significa dizer que a combinação repete exatamente o previsor original DEC.

4.5.3.2. Análise individual

A Figura 49 e a Tabela 87 exibem a evolução dos SMAPEs médios ao longo do horizonte de teste, considerando os previsores individuais e a melhor combinação tradicional obtida (de menor erro médio composto).

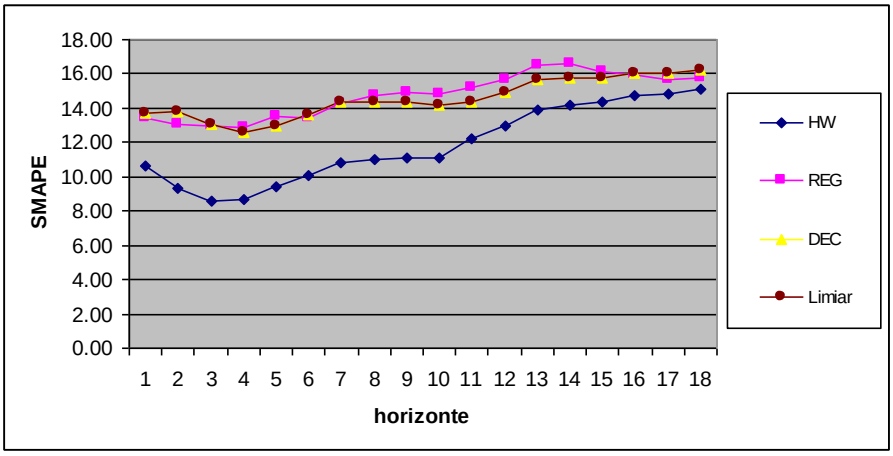

Figura 49 – Evolução dos SMAPEs médios fora da amostra (NN3).

Tabela 87 – Evolução dos SMAPEs médios fora da amostra (NN3)

h	*HW*	*REG*	*DEC*	*COMBINAÇÃO*
1	10.66	13.45	13.73	13.73
2	9.32	13.04	13.80	13.80
3	8.59	13.00	13.05	13.05
4	8.71	12.84	12.59	12.59
5	9.39	13.48	13.00	13.00
6	10.07	13.40	13.58	13.58
7	10.78	14.30	14.35	14.35
8	11.03	14.76	14.34	14.34
9	11.09	14.89	14.40	14.40
10	11.14	14.80	14.21	14.21
11	12.17	15.24	14.40	14.40
12	12.98	15.63	14.88	14.88
13	13.88	16.47	15.64	15.64
14	14.19	16.58	15.80	15.80
15	14.32	16.12	15.78	15.78
16	14.73	15.90	16.02	16.02
17	14.82	15.67	16.04	16.04
18	15.08	15.74	16.25	16.25

Melhor combinação limiar: *1c, AVG (=DEC).*

SMAPEs médios para as 11 séries da competição.

A Tabela 88 exibe as diferenças de desempenho médio tomadas período a período, fora da amostra, na ordem "erro médio do previsor individual (*benchmarking*) **menos** erro médio do previsor combinado". Como na seção 4.5.2.2, todos os testes de hipótese sugeridos para comparação de desempenhos foram executados; a Tabela 89 exibe os resultados obtidos.

Tabela 88 – Diferenças de desempenho médio individuais/combinação (NN3)

h	*HW*	*REG*	*DEC*
1	-3.07	-0.28	0.00
2	-4.48	-0.76	0.00
3	-4.46	-0.05	0.00
4	-3.88	0.25	0.00
5	-3.61	0.48	0.00
6	-3.51	-0.18	0.00
7	-3.57	-0.05	0.00
8	-3.31	0.42	0.00
9	-3.31	0.49	0.00
10	-3.07	0.59	0.00
11	-2.23	0.84	0.00
12	-1.90	0.75	0.00
13	-1.76	0.83	0.00
14	-1.61	0.78	0.00
15	-1.46	0.34	0.00
16	-1.29	-0.12	0.00
17	-1.22	-0.37	0.00
18	-1.17	-0.51	0.00
MEDIANA	**-3.07**	**0.30**	**0.00**

Tabela 89 – Testes de hipótese (NN3)

Combinação →	*1c, AVG (=DEC).*				
Teste t					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	-1	0.000	-3.28	-2.15	Normal
REG	0	0.120	-0.05	0.44	Normal
DEC	0	1.000	0.00	0.00	Normal
Teste de sinais					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	-1	0.000	-3.57	-1.61	Normal
REG	0	0.810	-0.18	0.59	Normal
DEC	0	1.000	0.00	0.00	Normal
Teste de Wilcoxon					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	-1	0.000	-3.36	-2.18	Normal
REG	0	0.140	-0.08	0.49	Normal
DEC	0	1.000	0.00	0.00	Normal
***srh0* (HW) = -3 *srh0* (REG) = 0 *srh0* (DEC) = 0**					

Para todos os testes são exibidos *valor-p* (*pvalue*), limites de confiança inferior (*inf*) e superior *(sup)* para a mediana observada e status do teste de normalidade Jarque-Bera (*JB*) para a distribuição estatística das diferenças de desempenho: *normal* ou *não*.

Com base nos valores observados (Tabela 89) para o indicador *srh0* (seção 4.4.2.2), pode-se chegar às conclusões da Tabela 90.

Tabela 90 – Conclusões para combinação limiar (NN3)

Benchmarking	*srh0*	*Conclusão*
HW	-3	A combinação é pior
REG	0	Indiferente
DEC	0	Indiferente
***srh0*+ = 0**		

4.5.4.
Combinações NEW

4.5.4.1.
Experimentos

Os experimentos com o sistema NEW foram realizados de maneira análoga à seção 4.4.4.1, focados apenas na geração de pesos **convexos**. Estes experimentos estão reunidos na Tabela 91, organizada de acordo com os previsores (**limiares**) sendo combinados, o método de geração de **pesos históricos** e a métrica de **desempenho aproximado**. Assim como na seção 4.5.3.1, os experimentos com 4 previsores limiares (prefixo "new2a") foram derivados dos modelos que apresentaram os melhores desempenhos no teste (HW e REG, Tabela 70).

Tabela 91 – Experimentos de combinação NEW

Experimento	*Previsores*	*Geração*	*Desempenho*
new1a_after_mad	HW+, HW-	AFTER	MAD
new1a_after_mse	HW+, HW-	AFTER	MSE
new1a_bg1_mad	HW+, HW-	BG1	MAD
new1a_bg1_mse	HW+, HW-	BG1	MSE
new1a_mqr_mad	HW+, HW-	MQR	MAD
new1a_mqr_mse	HW+, HW-	MQR	MSE
new1a_pool_mad	HW+, HW-	POOL	MAD
new1a_pool_mse	HW+, HW-	POOL	MSE
new2a_after_mad	HW+, HW-, REG+, REG-	AFTER	MAD
new2a_after_mse	HW+, HW-, REG+, REG-	AFTER	MSE
new2a_bg1_mad	HW+, HW-, REG+, REG-	BG1	MAD
new2a_bg1_mse	HW+, HW-, REG+, REG-	BG1	MSE
new2a_mqr_mad	HW+, HW-, REG+, REG-	MQR	MAD
new2a_mqr_mse	HW+, HW-, REG+, REG-	MQR	MSE
new2a_pool_mad	HW+, HW-, REG+, REG-	POOL	MAD
new2a_pool_mse	HW+, HW-, REG+, REG-	POOL	MSE

Neste estudo de caso, optou-se por usar **unicamente** referências do tipo SALY2. Foram dois os motivos para esta decisão: (i) as referências do tipo SALY2 apresentaram, em média, a menor **correlação conjunta** (seção 3.3) com os previsores originais (Tabela 93) e (ii) foram, de maneira bastante substancial,

as mais selecionadas nos experimentos anteriores (seção 4.4.4.1). A decisão por usar um único tipo de referência também reduziu de maneira significativa o número de experimentos. As Tabelas 92 e 93 apresentam, respectivamente, os desempenhos médios totais para cada tipo de referência e as **correlações de erro** (seção 2.2.2) dentro da amostra.

Tabela 92 – Desempenhos médios totais (referências)

Referência	*SMAPE Amostra*	*SMAPE Teste*
SALY1	20.39	13.94
SALY2	33.60	20.21
SALY3	21.86	17.60
DEC	14.10	16.25

SMAPEs médios para as 11 séries da competição.

Tabela 93 – Correlações de erro dentro da amostra

	SALY1	*SALY2*	*SALY3*	DEC	HW	REG
SALY1	**1.00**					
SALY2	0.71	**1.00**				
SALY3	0.95	0.77	**1.00**			
DEC	0.66	0.18	0.58	**1.00**		
HW	0.75	0.55	0.72	0.56	**1.00**	
REG	0.65	0.18	0.57	0.96	0.55	**1.00**

Valores médios para as 11 séries da competição.

Nos experimentos da Tabela 91, fixada a referência em SALY2, testaram-se os dois tipos de normalização possíveis (padrão e soma-1) e as duas janelas de tempo usadas anteriormente (seções 4.5.2.1 e 4.5.3.1): mínima e expansiva. Deste modo, o total de experimentos foi de **64** por série. A exemplo da seção 4.4.4.1, para cada experimento, selecionou-se a arquitetura de rede neural com menor erro de validação na **previsão combinada** (seção 3.9). As Tabelas 94 a 97 exibem os resultados (médios) selecionados, obtidos nos conjuntos de validação e teste; os valores em negrito indicam os experimentos com menor erro médio composto (SMAPE médio validação + SMAPE médio teste).

Tabela 94 – Desempenhos médios totais para *janela mínima & normalização padrão*

Experimento	*SMAPE Validação*	*SMAPE Teste*
new1a_after_mad	34.10	24.31
new1a_after_mse	30.79	24.44
new1a_bg1_mad	22.74	21.67
new1a_bg1_mse	19.31	17.60
new1a_mqr_mad	19.37	17.91
new1a_mqr_mse	19.92	18.95
new1a_pool_mad	19.11	17.79
new1a_pool_mse	18.32	16.62
new2a_after_mad	25.37	23.07
new2a_after_mse	23.05	21.17
new2a_bg1_mad	18.00	19.94
new2a_bg1_mse	17.87	16.61
new2a_mqr_mad	16.09	15.38
new2a_mqr_mse	16.94	16.04
new2a_pool_mad	16.93	14.52
new2a_pool_mse	16.90	16.09

SMAPEs médios para as 11 séries da competição.

O tipo de referência foi fixado em SALY2.

Tabela 95 – Desempenhos médios totais para *janela mínima & normalização soma-1*

Experimento	*SMAPE Validação*	*SMAPE Teste*
new1a_after_mad	30.76	24.11
new1a_after_mse	29.13	23.60
new1a_bg1_mad	22.76	20.20
new1a_bg1_mse	21.96	19.41
new1a_mqr_mad	19.93	17.18
new1a_mqr_mse	19.66	17.73
new1a_pool_mad	20.05	17.15
new1a_pool_mse	19.84	17.55
new2a_after_mad	23.03	20.63
new2a_after_mse	27.07	25.25
new2a_bg1_mad	19.38	17.89
new2a_bg1_mse	19.74	18.81
new2a_mqr_mad	16.71	15.50
new2a_mqr_mse	20.03	14.92
new2a_pool_mad	17.42	14.51
new2a_pool_mse	17.38	14.59

SMAPEs médios para as 11 séries da competição.

O tipo de referência foi fixado em SALY2.

Tabela 96 – Desempenhos médios totais para *janela expansiva & normalização padrão*

Experimento	*SMAPE Validação*	*SMAPE Teste*
new1a_after_mad	35.49	24.43
new1a_after_mse	45.81	29.09
new1a_bg1_mad	23.79	18.26
new1a_bg1_mse	25.69	21.99
new1a_mqr_mad	21.12	17.22
new1a_mqr_mse	23.45	21.56
new1a_pool_mad	19.71	24.07
new1a_pool_mse	24.56	15.95
new2a_after_mad	30.17	21.53
new2a_after_mse	29.20	23.75
new2a_bg1_mad	20.62	20.47
new2a_bg1_mse	17.85	15.29
new2a_mqr_mad	17.63	13.72
new2a_mqr_mse	**15.18**	**13.84**
new2a_pool_mad	17.11	13.30
new2a_pool_mse	16.34	14.17

SMAPEs médios para as 11 séries da competição.

O tipo de referência foi fixado em SALY2.

Tabela 97 – Desempenhos médios totais para *janela expansiva & normalização soma-1*

Experimento	*SMAPE Validação*	*SMAPE Teste*
new1a_after_mad	30.16	22.84
new1a_after_mse	29.35	29.22
new1a_bg1_mad	23.28	20.05
new1a_bg1_mse	23.22	26.52
new1a_mqr_mad	20.65	17.08
new1a_mqr_mse	20.22	19.72
new1a_pool_mad	20.65	20.81
new1a_pool_mse	20.54	20.48
new2a_after_mad	27.05	30.83
new2a_after_mse	28.29	26.12
new2a_bg1_mad	17.67	16.18
new2a_bg1_mse	17.50	15.89
new2a_mqr_mad	17.01	16.22
new2a_mqr_mse	15.92	13.83
new2_pool_mad	16.06	19.00
new2a_pool_mse	17.26	24.18

SMAPEs médios para as 11 séries da competição.
O tipo de referência foi fixado em SALY2.

4.5.4.2. Análise individual

A Tabela 98 e a Figura 50 exibem a evolução dos SMAPEs médios ao longo do horizonte de teste, considerando os previsores individuais e a melhor combinação tradicional obtida (de menor erro médio composto).

A Tabela 99 exibe as diferenças de desempenho médio tomadas período a período, fora da amostra, na ordem "erro médio do previsor individual (*benchmarking*) **menos** erro médio do previsor combinado". Como em seções anteriores, todos os testes de hipótese sugeridos para comparação destes desempenhos foram executados; a Tabela 100 exibe os resultados obtidos.

Tabela 98 – Evolução dos SMAPEs médios fora da amostra (NN3)

h	*HW*	*REG*	*DEC*	*COMBINAÇÃO*
1	10.66	13.45	13.73	8.40
2	9.32	13.04	13.80	8.72
3	8.59	13.00	13.05	9.23
4	8.71	12.84	12.59	9.69
5	9.39	13.48	13.00	10.65
6	10.07	13.40	13.58	10.67
7	10.78	14.30	14.35	11.76
8	11.03	14.76	14.34	12.10
9	11.09	14.89	14.40	12.10
10	11.14	14.80	14.21	12.03
11	12.17	15.24	14.40	12.53
12	12.98	15.63	14.88	13.23
13	13.88	16.47	15.64	14.25
14	14.19	16.58	15.80	14.42
15	14.32	16.12	15.78	14.01
16	14.73	15.90	16.02	13.86
17	14.82	15.67	16.04	13.76
18	15.08	15.74	16.25	13.84

Melhor combinação NEW: *new2a_mqr_mse, SALY2, janela expansiva & normalização padrão.*

SMAPEs médios para as 11 séries da competição.

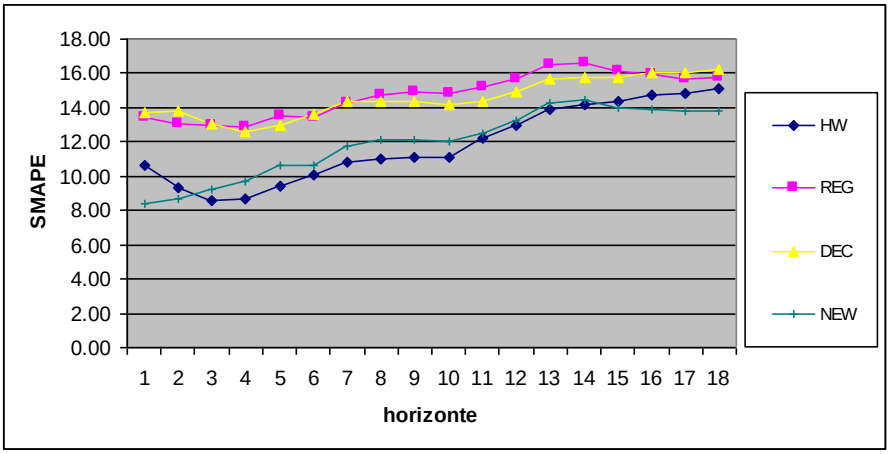

Figura 50 – Evolução dos SMAPEs médios fora da amostra (NN3).

Tabela 99 – Diferenças de desempenho médio individuais/combinação (NN3)

h	*HW*	*REG*	*DEC*
1	2.26	5.05	5.33
2	0.60	4.32	5.08
3	-0.64	3.77	3.82
4	-0.98	3.15	2.90
5	-1.26	2.83	2.35
6	-0.60	2.73	2.91
7	-0.98	2.54	2.59
8	-1.07	2.66	2.24
9	-1.01	2.79	2.30
10	-0.89	2.77	2.18
11	-0.36	2.71	1.87
12	-0.25	2.40	1.65
13	-0.37	2.22	1.39
14	-0.23	2.16	1.38
15	0.31	2.11	1.77
16	0.87	2.04	2.16
17	1.06	1.91	2.28
18	1.24	1.90	2.41
MEDIANA	**-0.37**	**2.68**	**2.29**

Tabela 100 – Testes de hipótese (NN3)

Combinação →	*new2a_mqr_mse, SALY2, janela expansiva & normalização padrão*				
Teste t					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	0	0.590	-0.61	0.36	Normal
REG	1	0.000	2.36	3.20	Não normal
DEC	1	0.000	2.03	3.14	Não normal
Teste de sinais					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	0	0.240	-0.98	0.60	Normal
REG	1	0.000	2.16	2.83	Não normal
DEC	1	0.000	1.87	2.90	Não normal
Teste de Wilcoxon					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
HW	0	0.400	-0.69	0.35	Normal
REG	1	0.000	2.35	3.21	Não normal
DEC	1	0.000	2.00	3.12	Não normal
***srh0* (HW) = 0** ***srh0* (REG) = 3** ***srh0* (DEC) = 3**					

Para todos os testes são exibidos *valor-p* (*pvalue*), limites de confiança inferior (*inf*) e superior *(sup)* para a mediana observada e status do teste de normalidade Jarque-Bera (*JB*) para a distribuição estatística das diferenças de desempenho: *normal* ou *não*.

Com base nos valores observados (Tabela 100) para o indicador *srh0* (seção 4.4.2.2), pode-se chegar às conclusões da Tabela 101. Vale destacar também a melhor configuração observada para o sistema NEW – previsores HW+, HW-, REG+ e REG- com referência SALY2 (fixada *a priori*), desempenhos aproximados MSE, geração de pesos históricos MQR, janela expansiva e normalização padrão. Esta configuração apresentou o melhor desempenho médio ao longo das 11 séries da competição NN3.

Tabela 101 – Conclusões para combinação NEW (NN3)

Benchmarking	*srh0*	*Conclusão*
HW	0	Indiferente
REG	3	A combinação é melhor
DEC	3	A combinação é melhor
***srh0*+ = 6**		

4.5.5.
Análise comparativa

4.5.5.1.
Comparação dirigida

De maneira análoga à seção 4.4.5.1, estabelece-se aqui uma comparação sem **viés** entre os métodos de combinação testados – **tradicional**, **limiar** e **NEW**. A ideia é comparar a melhor combinação NEW (explorada individualmente na seção 4.5.4.2) com as melhores combinações dos tipos limiar e tradicional, quando estas últimas estiverem restritas ao mesmo escopo – previsores e métodos de geração de pesos – definido para o sistema NEW. Neste caso, de acordo com a seção 4.5.4.1, assume-se para todos os métodos de combinação os previsores HW e REG e os métodos AFTER, BG1, MQR e AVG para geração de pesos.

A Figura 51 e a Tabela 102 exibem a evolução dos SMAPEs médios ao longo do horizonte de teste, considerando as combinações selecionadas para comparação dirigida (aquelas com menores erros médios compostos, limitadas ao escopo do NEW). As Tabelas 103 e 104 atualizam, respectivamente, as conclusões das Tabelas 81 e 90.

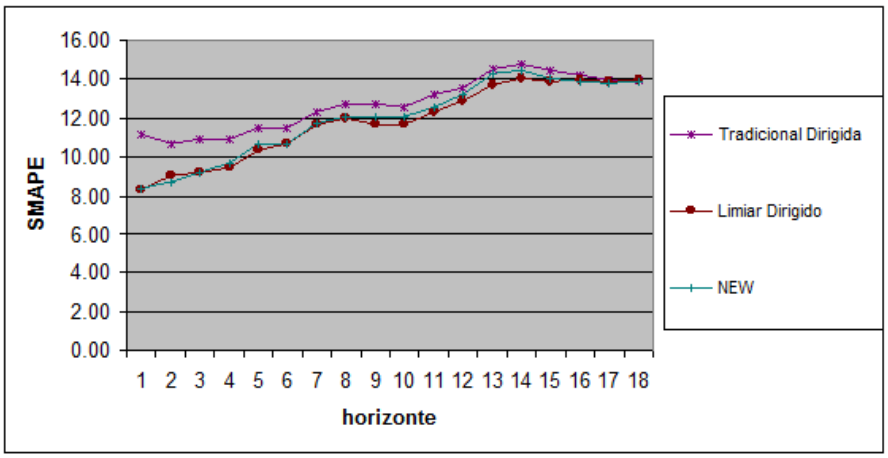

Figura 51 – Evolução dos SMAPEs médios fora da amostra (NN3).

A Tabela 105 exibe as diferenças de desempenho médio tomadas período a período, fora da amostra, na ordem "erro médio do previsor combinado (*benchmarking*) **menos** erro médio do previsor NEW". Como em seções

anteriores, todos os testes de hipótese sugeridos para comparação destes desempenhos foram executados; a Tabela 106 exibe os resultados obtidos.

Tabela 102 – Desempenhos médios para comparação dirigida (NN3)

h	*TRADICIONAL DIRIGIDA*	*LIMIAR DIRIGIDA*	*NEW*
1	11.12	8.29	8.40
2	10.69	9.05	8.72
3	10.90	9.22	9.23
4	10.89	9.47	9.69
5	11.49	10.34	10.65
6	11.46	10.69	10.67
7	12.31	11.62	11.76
8	12.68	11.98	12.10
9	12.74	11.64	12.10
10	12.58	11.65	12.03
11	13.19	12.29	12.53
12	13.51	12.87	13.23
13	14.53	13.74	14.25
14	14.76	14.00	14.42
15	14.41	13.85	14.01
16	14.19	13.93	13.86
17	13.92	13.84	13.76
18	13.95	13.98	13.84

Melhor combinação tradicional dirigida: *2a, geração MQR, janela expansiva & pesos estáticos;*
Melhor combinação limiar dirigida: *22a, AVG (= média HW-REG);*
Melhor combinação NEW: *new2a_mqr_mse, SALY2, janela expansiva & normalização padrão;*
SMAPEs médios para as 11 séries da competição.

Tabela 103 – Conclusões para combinação tradicional dirigida (NN3)

Benchmarking	*srh0*	*Conclusão*
HW	-3	A combinação é pior
REG	3	A combinação é melhor
DEC	3	A combinação é melhor
***srh0*+ = 6**		

Tabela 104 – Conclusões para combinação limiar dirigida (NN3)

Benchmarking	*srh0*	*Conclusão*
HW	0	Indiferente
REG	3	A combinação é melhor
DEC	3	A combinação é melhor
***srh0+* = 6**		

Tabela 105 – Diferenças de desempenho médio combinações/NEW (NN3)

h	*TRADICIONAL DIRIGIDA*	*LIMIAR DIRIGIDA*
1	2.72	-0.11
2	1.97	0.33
3	1.67	-0.01
4	1.20	-0.22
5	0.84	-0.31
6	0.79	0.02
7	0.55	-0.14
8	0.58	-0.12
9	0.64	-0.46
10	0.55	-0.38
11	0.66	-0.24
12	0.28	-0.36
13	0.28	-0.51
14	0.34	-0.42
15	0.40	-0.16
16	0.33	0.07
17	0.16	0.08
18	0.11	0.14
MEDIANA	**0.57**	**-0.15**

Tabela 106 – Testes de hipótese (NN3)

NEW →	*new2a_mqr_mse, SALY2, janela expansiva & normalização padrão*				
Teste t					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
TRAD. DIRIGIDA	1	0.000	0.44	1.13	Não normal
LIMIAR DIRIGIDA	-1	0.011	-0.27	-0.04	Normal
Teste de sinais					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
TRAD. DIRIGIDA	1	0.000	0.33	0.84	Não normal
LIMIAR DIRIGIDA	0	0.096	-0.36	0.02	Normal
Teste de Wilcoxon					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
TRAD. DIRIGIDA	1	0.000	0.42	1.13	Não normal
LIMIAR DIRIGIDA	-1	0.014	-0.29	-0.04	Normal
***srh0* (**TRAD. **DIRIGIDA) = 3** ***srh0* (LIMIAR DIRIGIDA) = -2**					

Para todos os testes são exibidos *valor-p* (*pvalue*), limites de confiança inferior (*inf*) e superior *(sup)* para a mediana observada e status do teste de normalidade Jarque-Bera (*JB*) para a distribuição estatística das diferenças de desempenho: *normal* ou *não*.

Com base nos valores observados (Tabela 106) para o indicador *srh0* (seção 4.4.2.2), pode-se chegar às conclusões da Tabela 107.

Tabela 107 – Conclusões para combinação NEW (NN3)

Benchmarking	*srh0*	*Conclusão*
TRAD.DIRIGIDA	3	O NEW é melhor
LIMIAR DIRIGIDA	-2	O NEW é pior

Por fim, a Tabela 108 exibe diferentes métricas de desempenho (calculadas fora da amostra) para os experimentos selecionados na comparação dirigida. Estas métricas são detalhadas **por série** nas Tabelas 109 a 111. Valores em negrito indicam superioridade do NEW (no acumulado 18 passos a frente).

Tabela 108 – Métricas de desempenho médio total na comparação dirigida (NN3)

Experimento	*SMAPE*	*RAE*	*UTHEIL*
TRAD. DIRIGIDA	13.95	0.38	0.33
LIMIAR DIRIGIDA	13.98	0.42	0.35
NEW	13.84	0.47	0.34

Valores médios para as 11 séries da competição.

Tabela 109 – Desempenhos SMAPE por série (NN3)

série	*HW*	*REG*	*DEC*	*TRADICIONAL*	*LIMIAR*	*NEW*
1	1.80	2.67	2.61	1.95	1.97	2.64
2	30.63	24.55	29.01	24.77	27.66	**22.05**
3	29.60	33.81	29.16	33.81	25.91	34.91
4	6.11	9.37	13.78	8.13	6.19	6.41
5	1.53	10.22	10.66	1.62	4.97	8.28
6	5.38	3.62	4.10	4.08	4.29	6.07
7	5.10	3.48	3.29	3.02	3.07	3.94
8	36.29	29.54	29.62	30.35	32.81	**25.83**
9	7.77	16.88	16.69	6.71	8.23	8.62
10	30.48	27.25	28.62	27.25	27.37	**23.02**
11	11.14	11.79	11.23	11.79	11.30	**10.49**
média	15.08	15.74	16.25	13.95	13.98	13.84

Melhor combinação tradicional dirigida: *2a, geração MQR, janela expansiva & pesos estáticos;*

Melhor combinação limiar dirigida: 22*a, AVG (= média HW-REG);*

Melhor combinação NEW: *new2a_mqr_mse, SALY2, janela expansiva & normalização padrão.*

Tabela 110 – Desempenhos RAE por série (NN3)

série	*HW*	*REG*	*DEC*	*TRADICIONAL*	*LIMIAR*	*NEW*
1	0.49	0.73	0.72	0.53	0.54	0.73
2	0.89	0.64	0.87	0.65	0.77	**0.56**
3	0.25	0.25	0.27	0.25	0.23	0.26
4	0.22	0.33	0.54	0.29	0.23	0.24
5	0.14	0.89	0.93	0.15	0.45	0.73
6	1.16	0.80	0.89	0.89	0.93	1.31
7	0.13	0.08	0.08	0.07	0.07	0.10
8	1.00	0.84	0.83	0.86	0.92	**0.74**
9	0.12	0.24	0.24	0.10	0.13	0.13
10	0.25	0.23	0.24	0.23	0.23	**0.20**
11	0.18	0.19	0.18	0.19	0.18	**0.17**
média	0.44	0.47	0.53	0.38	0.42	0.47

Melhor combinação tradicional dirigida: *2a, geração MQR, janela expansiva & pesos estáticos;*

Melhor combinação limiar dirigida: 22*a, AVG (= média HW-REG);*

Melhor combinação NEW: *new2a_mqr_mse, SALY2, janela expansiva & normalização padrão.*

Tabela 111 – Desempenhos UTHEIL por série (NN3)

série	*HW*	*REG*	*DEC*	*TRADICIONAL*	*LIMIAR*	*NEW*
1	0.04	0.07	0.06	0.05	0.05	0.07
2	0.49	0.37	0.48	0.37	0.43	**0.33**
3	0.19	0.22	0.25	0.22	0.18	0.20
4	0.18	0.19	0.29	0.18	0.17	0.19
5	0.04	0.24	0.24	0.05	0.13	0.19
6	0.14	0.09	0.11	0.11	0.11	0.14
7	0.20	0.14	0.13	0.14	0.14	0.17
8	0.91	0.76	0.74	0.78	0.83	**0.69**
9	0.53	0.93	0.95	0.42	0.49	0.53
10	1.04	0.90	1.01	0.90	0.91	**0.84**
11	0.47	0.45	0.47	0.45	0.45	**0.43**
média	0.39	0.40	0.43	0.33	0.35	0.34

Melhor combinação tradicional dirigida: *2a, geração MQR, janela expansiva & pesos estáticos;*

Melhor combinação limiar dirigida: 22*a, AVG (= média HW-REG);*

Melhor combinação NEW: *new2a_mqr_mse, SALY2, janela expansiva & normalização padrão.*

4.5.5.2. Comparação livre

Aqui, a exemplo da seção 4.4.5.2, os melhores resultados para cada método de combinação (explorados individualmente nas seções 4.5.2.2, 4.5.3.2 e 4.5.4.2) são comparados livremente entre si, sem qualquer restrição. A Tabela 112 e a Figura 52 exibem a evolução dos SMAPEs médios ao longo do horizonte de teste, considerando as combinações selecionadas para comparação livre (aquelas com menores erros médios compostos).

A Tabela 113 exibe as diferenças de desempenho médio tomadas período a período, fora da amostra, na ordem "erro médio do previsor combinado (*benchmarking*) **menos** erro médio do previsor NEW". Como em seções anteriores, todos os testes de hipótese sugeridos para comparação destes desempenhos foram executados; a Tabela 114 exibe os resultados obtidos.

Tabela 112 – Desempenhos médios para comparação livre (NN3)

h	*TRADICIONAL*	*LIMIAR*	*NEW*
1	9.47	13.73	8.40
2	9.96	13.80	8.72
3	9.39	13.05	9.23
4	8.87	12.59	9.69
5	9.18	13.00	10.65
6	9.70	13.58	10.67
7	10.44	14.35	11.76
8	10.65	14.34	12.10
9	10.79	14.40	12.10
10	10.80	14.21	12.03
11	11.50	14.40	12.53
12	12.12	14.88	13.23
13	13.16	15.64	14.25
14	13.48	15.80	14.42
15	13.41	15.78	14.01
16	13.63	16.02	13.86
17	13.55	16.04	13.76
18	13.71	16.25	13.84

Melhor combinação tradicional: *3a, geração AFTER, janela expansiva & pesos estáticos;*

Melhor combinação limiar: *1c, AVG (= DEC);*

Melhor combinação NEW: *new2a_mqr_mse, SALY2, janela expansiva & normalização padrão;*

SMAPEs médios para as 11 séries da competição.

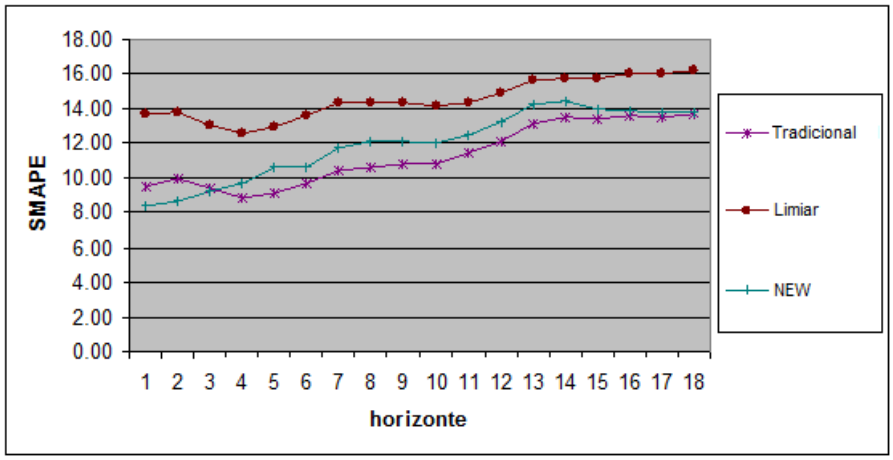

Figura 52 – Evolução dos SMAPEs médios fora da amostra (NN3).

Tabela 113 – Diferenças de desempenho médio combinações/NEW (NN3)

h	*TRADICIONAL*	*LIMIAR*
1	1.07	5.33
2	1.24	5.08
3	0.16	3.82
4	-0.82	2.90
5	-1.47	2.35
6	-0.97	2.91
7	-1.32	2.59
8	-1.45	2.24
9	-1.31	2.30
10	-1.23	2.18
11	-1.03	1.87
12	-1.11	1.65
13	-1.09	1.39
14	-0.94	1.38
15	-0.60	1.77
16	-0.23	2.16
17	-0.21	2.28
18	-0.13	2.41
MEDIANA	**-0.96**	**2.29**

Tabela 114 – Testes de hipótese (NN3)

NEW →	*new2a_mqr_mse, SALY2, janela expansiva & normalização padrão*				
Teste t					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
TRADICIONAL	-1	0.000	-1.04	-0.23	Não normal
LIMIAR	1	0.000	2.03	3.14	Não normal
Teste de sinais					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
TRADICIONAL	-1	0.010	-1.23	-0.21	Não normal
LIMIAR	1	0.000	1.87	2.90	Não normal
Teste de Wilcoxon					
Benchmarking	H_0	*pvalue*	*inf*	*sup*	*JB*
TRADICIONAL	1	0.010	-1.13	-0.18	Não normal
LIMIAR	1	0.000	2.00	3.12	Não normal
***srh0* (**TRADICIONAL**) = -3** ***srh0* (LIMIAR) = 3**					

Para todos os testes são exibidos *valor-p* (*pvalue*), limites de confiança inferior (*inf*) e superior *(sup)* para a mediana observada e status do teste de normalidade Jarque-Bera (*JB*) para a distribuição estatística das diferenças de desempenho: *normal* ou *não*.

Com base nos valores observados (Tabela 114) para o indicador *srh0* (seção 4.4.2.2), pode-se chegar às conclusões da Tabela 115.

Tabela 115 – Conclusões para combinação NEW (NN3)

Benchmarking	*srh0*	*Conclusão*
TRADICIONAL	-3	O NEW é pior
LIMIAR	3	O NEW é melhor

Por fim, a Tabela 116 exibe diferentes métricas de desempenho (calculadas fora da amostra) para os experimentos selecionados na comparação dirigida. Estas métricas são detalhadas **por série** nas Tabelas 117 e 119; valores em negrito indicam superioridade do NEW (no acumulado18 passos a frente).

Tabela 116 – Métricas de desempenho médio total na comparação livre (NN3)

Experimento	*SMAPE*	*RAE*	*UTHEIL*
TRADICIONAL	13.71	0.40	0.35
LIMIAR	16.25	0.53	0.43
NEW	13.84	0.47	0.34

Valores médios para as 11 séries da competição.

Tabela 117 – Desempenhos SMAPE por série (NN3)

série	*HW*	*REG*	*DEC*	*TRADICIONAL*	*LIMIAR*	*NEW*
1	1.80	2.67	2.61	2.43	2.61	2.64
2	30.63	24.55	29.01	24.56	29.01	**22.05**
3	29.60	33.81	29.16	28.52	29.16	34.91
4	6.11	9.37	13.78	9.39	13.78	6.41
5	1.53	10.22	10.66	1.53	10.67	8.28
6	5.38	3.62	4.10	4.06	4.10	6.07
7	5.10	3.48	3.29	3.29	3.29	3.94
8	36.29	29.54	29.62	29.53	29.62	**25.83**
9	7.77	16.88	16.69	7.76	16.69	8.62
10	30.48	27.25	28.62	28.48	28.62	**23.02**
11	11.14	11.79	11.23	11.27	11.23	**10.49**
média	15.08	15.74	16.25	13.71	16.25	13.84

Melhor combinação tradicional: *3a, geração AFTER, janela expansiva & pesos estáticos;*

Melhor combinação limiar: *1c, AVG (= DEC);*

Melhor combinação NEW: *new2a_mqr_mse, SALY2, janela expansiva & normalização padrão.*

Tabela 118 – Desempenhos RAE por série (NN3)

série	*HW*	*REG*	*DEC*	*TRADICIONAL*	*LIMIAR*	*NEW*
1	0.49	0.73	0.72	0.67	0.72	0.73
2	0.89	0.64	0.87	0.64	0.87	**0.56**
3	0.25	0.25	0.27	0.27	0.27	0.26
4	0.22	0.33	0.54	0.33	0.54	0.24
5	0.14	0.89	0.93	0.14	0.93	0.73
6	1.16	0.80	0.89	0.89	0.89	1.31
7	0.13	0.08	0.08	0.08	0.08	0.10
8	1.00	0.84	0.83	0.84	0.83	**0.74**
9	0.12	0.24	0.24	0.12	0.24	0.13
10	0.25	0.23	0.24	0.24	0.24	**0.20**
11	0.18	0.19	0.18	0.18	0.18	**0.17**
média	0.44	0.47	0.53	0.40	0.53	0.47

Melhor combinação tradicional: *3a, geração AFTER, janela expansiva & pesos estáticos;*

Melhor combinação limiar: *1c, AVG (= DEC);*

Melhor combinação NEW: *new2a_mqr_mse, SALY2, janela expansiva & normalização padrão.*

Tabela 119 – Desempenhos UTHEIL por série (NN3)

série	*HW*	*REG*	*DEC*	*TRADICIONAL*	*LIMIAR*	*NEW*
1	0.04	0.07	0.06	0.06	0.06	0.07
2	0.49	0.37	0.48	0.37	0.48	**0.33**
3	0.19	0.22	0.25	0.25	0.25	0.20
4	0.18	0.19	0.29	0.19	0.29	0.19
5	0.04	0.24	0.24	0.04	0.24	0.19
6	0.14	0.09	0.11	0.11	0.11	0.14
7	0.20	0.14	0.13	0.13	0.13	0.17
8	0.91	0.76	0.74	0.75	0.74	**0.69**
9	0.53	0.93	0.95	0.53	0.95	**0.53**
10	1.04	0.90	1.01	1.00	1.01	**0.84**
11	0.47	0.45	0.47	0.47	0.47	**0.43**
média	0.39	0.40	0.43	0.35	0.43	0.34

Melhor combinação tradicional: *3a, geração AFTER, janela expansiva & pesos estáticos;*

Melhor combinação limiar: *1c, AVG (= DEC);*

Melhor combinação NEW: *new2a_mqr_mse, SALY2, janela expansiva & normalização padrão.*

4.5.5.3. Tempo de processamento

Para estimar o tempo total de processamento consumido neste estudo de caso, pode-se utilizar os mesmos tempos médios da seção 4.4.5.3 (Tabela 69). No caso das combinações **limiares**, houve 64 experimentos por série, sendo 48 do tipo Limiar1 e 16 do tipo Limiar2 (seção 4.5.3.1); desta forma, para as 11 séries da competição NN3, admitindo-se execução sequencial, tem-se um tempo total estimado de (aproximadamente) **26 minutos**. No caso das combinações **NEW**, houve também 64 experimentos por série, mas com 32 do tipo NEW1 e 32 do tipo NEW2 (seção 4.5.4.1); desta forma, o tempo total estimado varia entre (aproximadamente) **3hs** (cenário otimista) e **34 dias** (cenário conservador).

4.5.5.4. Resumo

Considerando o critério de avaliação período a período, realizado pelo teste estatístico da variável **diferença de desempenho**, os resultados da comparação **dirigida** (seção 4.5.5.1) deixam o sistema NEW em vantagem sobre o método tradicional, mas em desvantagem em relação ao método limiar (Tabela 107). Contudo, por dois outros critérios, pode-se observar vantagens: (i) o NEW apresentou, na comparação dirigida, o menor SMAPE médio total fora da amostra (acumulado em 18 meses) – **13.84** (Tabela 108) – e (ii) foi o método que prevaleceu no maior número de séries da competição: 4 em 11 (Tabela 109). Em particular, a análise das Tabelas 109 a 111 mostra que o sistema NEW foi particularmente bem sucedido nas séries de números **2**, **8**, **10** e **11**[18] (Apêndice C). Ainda na comparação dirigida, o indicador ***srh0+*** (seção 4.4.2.2) mostra que os métodos de combinação testados, quando comparados com os previsores individuais, são geralmente melhores: o *srh0+* vale **6** (em um máximo de **9)** para todos os métodos (Tabelas 103, 104 e 101).

[18] A análise desse resultado é uma das sugestões para estudos futuros.

Na comparação **livre** (seção 4.5.5.2), a análise do indicador *srh0+* – com valores de **6** para o modelo NEW, **0** para a combinação limiar e **9** para a combinação tradicional (Tabelas 81, 90 e 101) – indica vantagens no uso do sistema NEW e da combinação tradicional (principalmente).

Na comparação direta entre os métodos de combinação, o NEW permaneceu com desempenho intermediário, mas as posições entre os métodos tradicional e limiar se alternaram (Tabela 115): o método limiar, quando exposto a todos os previsores, sofreu uma perda de **2.27** pp em seu desempenho, enquanto o método tradicional melhorou em **0.24** pp (Tabelas 108 e 116). Neste cenário de comparação livre, o sistema NEW continuou prevalecendo no maior número de séries (Tabelas 117 a 119).

O melhor método de combinação tradicional observado neste trabalho (Tabela 112) atingiu um SMAPE médio total de **13.71 (**Tabela 116). Uma curiosidade[19] neste resultado é o fato da geração de pesos ser estática: para cada série, repete-se, ao longo de todo o horizonte de previsão, os últimos pesos calculados no histórico. Apesar do bom desempenho, a comparação do melhor resultado tradicional com o melhor resultado NEW pode estar enviesada, na medida em que a combinação tradicional utiliza um previsor fora do escopo dirigido: DEC. O possível viés na comparação com o método tradicional livre e a própria limitação de escopo nos experimentos (referência fixa e menor número de previsores componentes) são fatores que podem ser elencados para reforçar a utilidade do sistema NEW e incentivar estudos futuros. Além disso, esta metodologia pode ser sempre desejável por sua capacidade de encapsular a complexidade do processo de combinar de previsores.

4.5.6. Resultados da competição

Os resultados finais da competição NN3 estão integralmente publicados na *internet* (NN3, 2011).

[19] Apesar de curioso, esse fato não chega a ser surpreendente. Como mencionado na seção 1.4, não há garantia de que a geração dinâmica seja a melhor opção sempre.

A Figura 53 exibe os 10 primeiros colocados para a competição reduzida (seção 4.5). A métrica usada no *ranking* é o SMAPE médio para as 11 séries, considerando conjuntos de teste com 18 meses de dados.

Levando em conta apenas a comparação dirigida (seção 4.5.5.1), o menor SMAPE médio relatado neste trabalho foi de **13.84,** obtido pelo NEW (Tabela 108); este resultado garantiria ao método a **segunda** colocação na competição NN3. Por outro lado, considerando-se a comparação livre (seção 4.5.5.2), o menor SMAPE médio foi de **13.71,** obtido pela combinação tradicional (Tabela 116); este resultado também garantiria a **segunda** colocação.

Rank on SMAPE	Participant	SMAPE
	CI Benchmark - Theta AI (Nikolopoulos)	13,07%
	Stat. Benchmark - Autobox (Reily)	13,49%
	Stat. Benchmark - ForecastPro (Stellwagen)	13,52%
1	Yan	13,68%
	Stat. Benchmark - Theta (Nikolopoulos)	13,70%
2	Ilies, Jäger, Kosuchinas, Rincon, Sakenas, Vaskevcius	14,26%
3	Chen, Yao	14,46%
4	Yousefi, Miromeni, Lucas	14,49%
5	Ahmed, Atiya, Gayar, El-Shishiny	14,52%
6	Flores, Anaya, Ramirez, Morales	15,00%
7	Adeodato, Vasconcelos, Arnaud, Chunha, Monteiro	15,10%
	Stat. Contender - Wildi	15,32%
8	Luna, Soares, Ballini	15,35%
9	Theodosiou, Swamy	16,19%
10	Hwang, Song, Kasabov	16,31%

Figura 53 – 10 primeiros colocados da competição NN3 reduzida, considerando o SMAPE médio. Os resultados em vermelho ou azul participaram como *benchmarking.*

É importante lembrar que os métodos de combinação testados aqui não consideraram nenhum tipo de **pré-processamento** estatístico (e.g. eliminação de *outliers*) sobre as séries originais. Este pode não ter sido o caso de alguns dos métodos apresentados no *ranking* final da competição.