Avaliação da utilidade da aplicação preliminar de análise de componentes principais em um grande conjunto de dados quando se visa construir modelos de regressão

Resumo

A análise multivarida se contitui em um conjunto de técnicas para se estimar parâmetros de um conjunto de variáveis de uma mesma unidade experimental. Dentro deste conjunto a análise de regressão é um método para se modelar a dependência entre variáveis em estudo e a análise de componentes principais (ACP), uma das técnicas para se modelar sua independência.

Na medida em que a ACP é uma técnica que permite se substituir um conjunto de variáveis que possuam independência estatística e correlação significativa, por um conjunto de novas variáveis não-correlacionadas, através de sua combinação linear na forma original, questiona-se se não seria útil sua pré-aplicação em situações em que se trabalhe com um grande conjunto de dados onde se tenha o intuito de construir modelos de regressão. Tal questão surge na medida em que sabe-se que algumas das vantagens de uma ACP estão em diminuir a dimensionalidade das variáveis envolvidas no processo de análise, facilitando a interpretação dos resultados mantendo um aceitável grau de explicação do fenômeno em estudo devido reduzida perda de informações.

Buscando iluminar a discussão em torno deste tema, este artigo se popõe a construir modelos de regressão baseando-se na pré-aplicação da análise de componentes principais, combinar suas predições, e compará-la com as predições construídas a partir de um modelo de regressão estimado sem a pré-aplicação da técnica de ACP.

As medidas utilizadas como critério de comparação para avaliação da utilidade de uma pré-aplicação da ACP na construção de modelos de regressão quando se trabalha com grandes conjuntos de dados, serão o Erro Percentual Absoluto Médio (EPAM) e a estatística U de Theil.

Palavras chave: Análise de Componentes Principais; Modelos de Regressão Dinâmica; Combinação de Previsões.

1. Introdução

Segundo Lopes (2001), é possível encontrar componentes, que serão combinações lineares das variáveis originais ordenados decrescentemente em termos de variância, a partir do estudo de "n" observações distribuídas em "x" variáveis correlacionadas de um sistema. Em termos de detecção de erros, segundo ele, utilizar variáveis provenientes da ACP pode ser tão eficiente quanto utilizar os próprios dados originais, devido ao fato de que os vetores estarão representando o comportamento conjunto destas variáveis, o que permite que as "x" variáveis originais sejam substituídas por estes componentes sem muita perda de informação.

Baseando-se nisto, procura-se discutir a questão de ser ou não útil a pré-aplicação da ACP em situações onde se trabalhe com um grande conjunto de dados e tenha o objetivo de construir modelos de regressão, tornando possível a redução nas discrepâncias. Para isto, neste artigo se construirá modelos de regressão baseando-se na pré-aplicação da análise de componentes principais. Posteriormente se combinará tais predições e as comparará com as predições construídas a partir de um modelo de regressão estimado sem a pré-aplicação da técnica de

ACP. Esta metodologia pode ser mais bem compreendida observando a Figura 1.

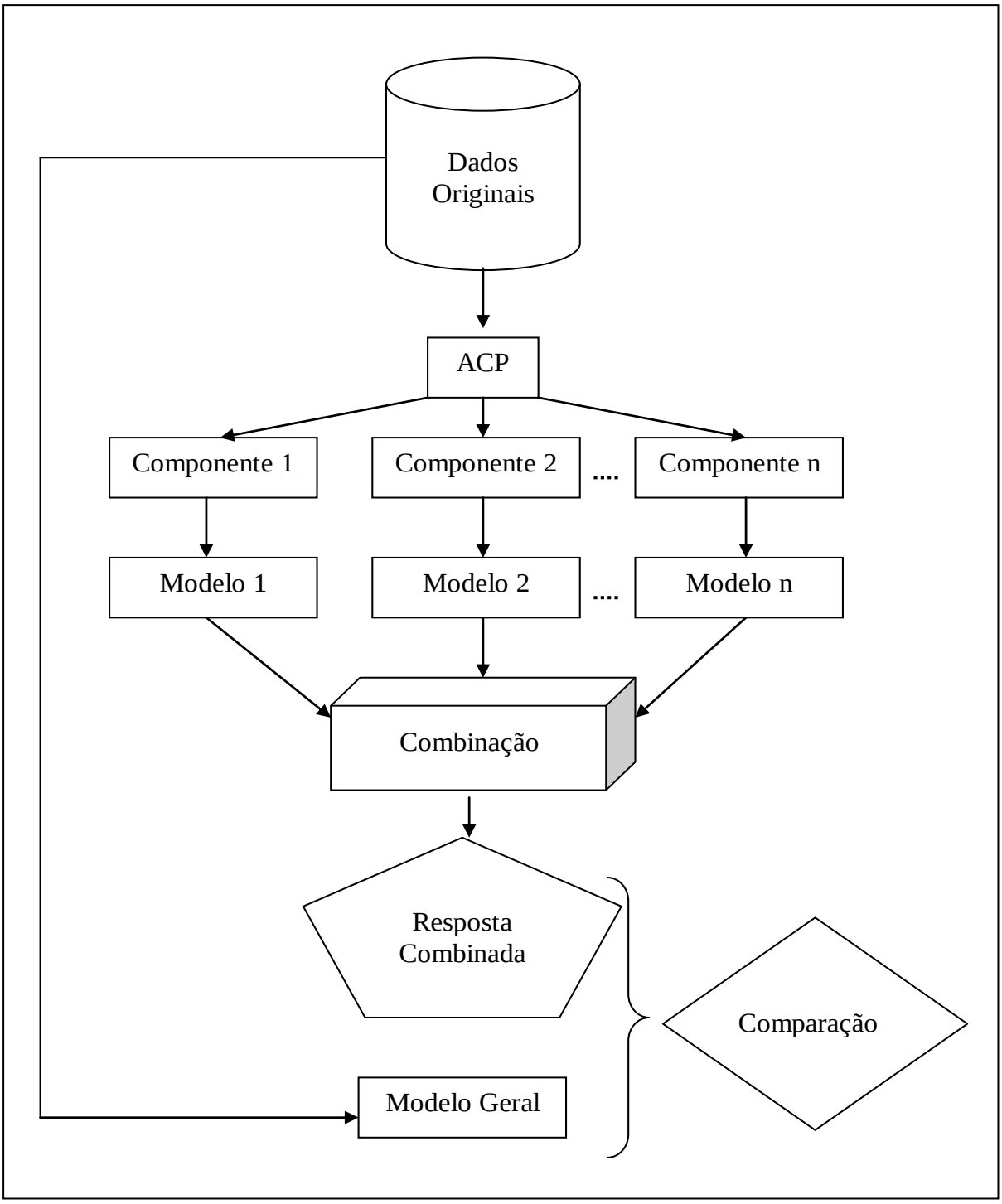

Figura 1 – Fluxograma do processo utilizado para comparação das metodologias

Para a ilustração da metodologia a ser aplicada, foram selecionados os dados referentes à emissão de cheques sem fundo (CSF), mais especificamente o percentual da segunda devolução em cada 1.000 cheques compensados. Também foram coletadas outras variáveis que estivessem correlacionadas com a mesma, em função da conjuntura macro-econômica do país, de forma a permitir relacionar a qual destes indicadores macroecômicos e de atividade econômica estaria ligada a variável predita.

A relação das variáveis utilizadas no estudo, visualizáveis na Tabela 1, foram coletadas com

periodicidade mensal de agosto de 1995 a abril de 2005.

Variável	Acrônimo
Taxa de Câmbio	TC
Termos de Troca	TT
Arrecadação Bruta da Receita Federal	ARF
Dívida Interna Líquida do Setor Público	DI
Taxa de Juros Over-Selic	TJ
Índice Geral de Preços de Mercado	IGP-M
Utilização da Capacidade Industrial Instalada	UCI
Produção Física de Alimentos	PAF
Salário Mínimo Real	SM
Emprego Formal na Atividade Agropecuária, Extrativa Vegetal e de Caça e Pesca	EA-EV-CP
Emprego Formal na Atividade Extrativa Mineral	EEM
Emprego Formal na Indústria de Transformação	EIT
Emprego Formal na Indústria de Minerais Não-metálicos	EMNM
Emprego Formal na Metalurgia	EM
Emprego Formal na Indústria Mecânica	EMC
Emprego Formal na Indústria de Materiais Elétricos e de Comunicação	EMEC
Emprego Formal na Indústria de Materiais de Transporte	EMT
Emprego Formal na Indústria de Mobiliário	EMOB
Emprego Formal na Indústria Editorial e Gráfica	EEG
Emprego Formal na Indústria de Fumo, Couros e Produtos Similares	EFC
Emprego Formal na Indústria Química e Produtos Farmacêuticos	EQPF
Emprego Formal na Indústria Têxtil, Vestuário e Artefatos de Couro	ET
Emprego Formal na Indúutria de Calçados	EC
Emprego Formal na Indústria de Produtos Alimentares e Bebidas	EPAB
Emprego Formal nos Serviços Industriais de Utilidade Pública	ESIUP
Emprego Formal na Construção Civil	ECC
Emprego Formal no Comércio	ECOM
Emprego Formal nos Serviços	ES
Emprego Formal na Administração Pública Direta e Autárquica	EAPA

Fonte: Instituto de Pesquisa Econômica Aplicada (www.ipeadata.com.br) e Banco Central do Brasil (www4.bcb.gov.br/pec/series/port).

Tabela 1 – Variáveis explicativas consideradas no estudo

Como observado anteriormente, na ACP foi considerada a variável CSF como sendo a suplementar, e a mesma como dependente nos modelos de regressão dinâmica. Já a avaliação das predições de cada modelo e deles combinados é baseada no resultado do cálculo do EPAM (Erro Percentual Absoluto Médio) e do U de Theil (ou coeficiente de desigualdade), onde o EPAM representa a percentagem do erro médio da predição, e o U de Theil mede a desigualdade percentual entre os valores preditos e os verificados.

Cada um dos modelos utilizados, tanto os individuais como o geral foram desenvolvidos no *software* aplicativo PcGive 10 e a combinação realizada no *software* aplicativo Excel, onde se utilizou a ferramenta Solver para encontrar os respectivos pesos de cada modelo de forma que

minimizassem seu EPAM.

2. Metodologia

Ao realizar-se a ACP das variáveis padronizadas se utilizou a matriz de correlação, o que exigiu que o critério de seleção dos componentes fosse o de Kaiser, ou seja, incluir-se-iam na análise apenas aqueles que tivessem os autovalores maiores ou iguais a um. Este critério permite afirmar que o conjunto de variáveis dos respectivos componentes explica o fenômeno em estudo com maior precisão do que apenas uma das variáveis individualmente.

Esta situação pode ser observada na Tabela 2 e ratificada pela utilização do critério de do gráfico de Cattel, observável na Figura 2, onde, segundo Lopes (2001), o número de componentes deve ser determinado pelo corte, na curva dos pesos dos autovalores, no ponto onde ocorre sua estabilização.

Componentes	Autovalores	Variação Total (%)	Autovalor Cumulativo	% Cumulativo
1	16,23815	55,99362	16,23815	55,9936
2	9,10397	31,39299	25,34212	87,3866
3	1,01396	3,49642	26,35608	90,8830
4	0,78168	2,69546	27,13776	93,5785
5	0,54721	1,88692	27,68497	95,4654
6	0,31982	1,10285	28,00479	96,5682
7	0,22954	0,79151	28,23433	97,3598
8	0,20484	0,70633	28,43917	98,0661
9	0,12350	0,42588	28,56267	98,4920
10	0,09601	0,33108	28,65868	98,8230
11	0,08455	0,29155	28,74323	99,1146
12	0,06764	0,23322	28,81087	99,3478
13	0,04478	0,15443	28,85565	99,5023
14	0,03252	0,11215	28,88818	99,6144
15	0,02787	0,09612	28,91605	99,7105
16	0,01871	0,06452	28,93476	99,7750
17	0,01751	0,06038	28,95227	99,8354
18	0,01363	0,04699	28,96590	99,8824
19	0,00923	0,03183	28,97513	99,9142
20	0,00848	0,02925	28,98361	99,9435
21	0,00431	0,01485	28,98792	99,9583
22	0,00370	0,01276	28,99162	99,9711
23	0,00281	0,00970	28,99443	99,9808
24	0,00164	0,00565	28,99607	99,9865
25	0,00141	0,00487	28,99748	99,9913
26	0,00117	0,00403	28,99865	99,9953
27	0,00086	0,00297	28,99951	99,9983
28	0,00043	0,00147	28,99994	99,9998
29	0,00006	0,00021	29,00000	100,0000

Fonte: Aplicativo Excel.

Tabela 2 – Autovalores baseados na matriz de correlação

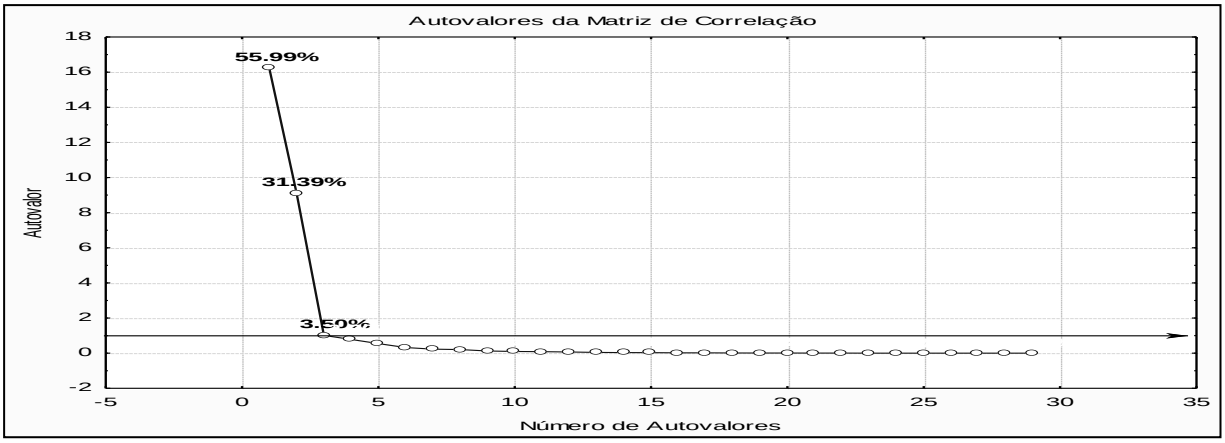

Figura 2 – Gráfico dos pesos dos autovalores

É importante mencionar, que para cada um dos componentes selecionados, consideraram-se apenas aquelas variáveis definidas pela combinação linear que apresentassem contribuição acima da média no componente, visando-se evitar a situação de se construir modelos idênticos, já que na ACP, todas as variáveis são consideradas em cada um dos componentes individualmente. A contribuição de cada variável dentro de cada componente gerada pela sua combinação linear pode ser visualizada na Tabela 3.

Variável	Componente 1	Componente 2	Componente 3
TC	0,021966	0,048618	0,042774
TT	0,018278	0,027182	0,182858
ARF	0,045224	0,020774	0,002020
DI	0,042564	0,030950	0,000072
TJ	0,006775	0,055762	0,124066
IGP-M	0,049868	0,017100	0,000583
UCI	0,030646	0,003311	0,148413
PAF	0,027780	0,019684	0,156445
SM	0,032899	0,030730	0,024574
EA-EV-CP	0,003718	0,063427	0,164110
EEM	0,040124	0,034237	0,005556
EIT	0,036409	0,044438	0,000237
EMNM	0,050402	0,000937	0,001274
EM	0,039574	0,037154	0,003033
EMC	0,045331	0,025714	0,004387
EMEC	0,008250	0,084725	0,004766
EMT	0,014965	0,080252	0,003571
EMOB	0,059630	0,000146	0,003105
EEG	0,014290	0,078153	0,001979
EFC	0,042446	0,024047	0,021029
EQPF	0,051970	0,015618	0,001662
ET	0,038199	0,029873	0,022639
EC	0,059451	0,000148	0,001380
EPAB	0,011572	0,076070	0,014746
ESIUP	0,010047	0,088526	0,001214
ECC	0,027028	0,053472	0,039911
ECOM	0,058491	0,001088	0,000862
ES	0,059680	0,000716	0,000202
EAPA	0,052423	0,007148	0,022536
MÉDIA	**0,034483**	**0,034483**	**0,034483**

Fonte: Aplicativo Excel.

Tabela 3 – Contribuição das variáveis nos componentes baseada na correlação

Segundo o critério de Kaiser e do gráfico de Cattel, observado na Figura 2, considerou-se os três primeiros componentes como sendo os principais, os quais explicam aproximadamente 90,88% do comportamento das variáveis originais, permitindo que as mesmas sejam por eles substituídas sem muita perda de informações. Consequentemente houve a necessidade de se construir igualmente três modelos de regressão dinâmica, um para cada um dos componentes considerados, cada qual com sua particularidade.

Para cada um dos modelos construídos, considerou-se como variável dependente a diferença dos logaritmos da variável CSF (DLCSF), visando respeitar as condições de estacionaridade e normalidade da série. Condições estas não exigíveis na ACP, porém fundamentais na aplicabilidade da técnica de mínimos quadrados ordinários utilizada para estimar os parâmetros dos modelos de regressão, na medida em que contribuem para que os resíduos dos modelos finais se apresentem normalmente distribuídos, com média zero e variância constante. Lembra-se também que esta transformação foi considerada nas demais variáveis explicativas já mencionadas, exceto para a variável EEG, onde se considerou a primeira diferença de sua padronização ($DEEG_p$).

No primeiro modelo, correspondente ao primeiro componente selecionado, identificado no trabalho como MOD-01, foram consideradas cinco defasagens da variável dependente e de cada uma de suas 16 independentes (ARF, DI, IGP-M, EEM, EIT, EMNM, EM, EMC, EMOB, EFC, EQPF, ET, EC, ECOM, ES e EAPA) além de uma constante. Este número de defasagens foi determinado, tendo em vista as 123 observações consideradas, pois este tamanho de amostra não permitiria um número maior de defasagens.

O modelo contou inicialmente com 102 parâmetros, que, após a realização das estimativas, reduziram-se a 29, conforme demonstrado a seguir. Este resultado apresentou um R^2 de 80,87% e um R^2_{ajust} de 74,92%.

$$
\begin{aligned}
DLCSF_t = {} & +0,097 - 0,810\,DLCSF_{t-1} - 0,444\,DLCSF_{t-2} - 0,248\,DLCSF_{t-3} - 0,188\,DLCSF_{t-4} \\
& -0,324\,DLCSF_{t-5} - 0,957\,DLDI_{t-4} + 3,013\,DLIGP-M_{t-3} - 7,154\,DLEEM_{t-1} - 6,980\,DLEEM_{t-3} \\
& +6,205\,DLEIT_{t-1} - 4,051\,DLEIT_{t-4} + 7,267\,DLEMNM_{t-1} + 11,071\,DLEMNM_{t-2} - 7,553\,DLEMNM_{t-3} \\
& +6,189\,DLEM_t + 13,804\,DLEM_{t-3} + 16,679\,DLEM_{t-5} + 3,966\,DLEMC_{t-1} - 6,261\,DLEMOB_{t-2} \\
& -6,564\,DLEMOB_{t-5} - 1,873\,DLEFC_{t-1} - 2,391\,DLEFC_{t-3} - 11,854\,DLEQPF_{t-5} - 6,919\,DLET_t \\
& +5,861\,DLET_{t-1} - 4,362\,DLEC_{t-1} - 12,633\,DLECOM_{t-1} - 16,365\,DLES_{t-3}
\end{aligned}
$$

O segundo modelo construído (MOD-02), representando o segundo componente selecionado, e como contou com um número reduzido de variáveis em relação ao primeiro (11), foi possível considerar até oito defasagens de cada uma das variáveis (CSF, TC, TJ, EA-EV-CP, EIT, EM, EMEC, EMT, EEG, EPAB, ESIUP, ECC), além de um parâmetro constante. O MOD-02 partiu de um modelo geral com 108 parâmetros para 116 observações, chegando-se a um modelo específico com apenas 15, como demonstrado abaixo. Este modelo apresentou um R^2 de 65,96% e um R^2_{ajust} de 60,60%.

$$
\begin{aligned}
DLCSF_t = {} & -0,663\,DLCSF_{t-1} - 0,221\,DLCSF_{t-2} + 0,229\,DLCSF_{t-6} - 0,549\,DLTC_{t-4} \\
& +0,519\,DLTC_{t-5} - 0,805\,DLEA-EV-CP_{t-1} + 7,331\,DLEIT_{t-4} + 10,310\,DLEM_{t-1} - 2,572\,DLEMT_{t-3} \\
& -2,566\,DLEPAB_{t-3} - 3,313\,DLEPAB_{t-4} - 5,868\,DLESIUP_{t-6} + 4,234\,DLESIUP_{t-7} + 2,863\,DLECC_{t-6} \\
& -0,374\,DEEGp_{t-1}
\end{aligned}
$$

Já no terceiro modelo construído (MOD-03), teve como variáveis explicativas apenas TC, TT, TJ, UCI, PFA, EA-EV-CP e EEC, o que permitiu que se utilizassem 12 defasagens para cada

uma das variáveis. O modelo geral teve 108 parâmetros para 116 observações, e após especificá-lo, encontrou-se um modelo com apenas 38 parâmetros, como demonstrado a seguir. Esta equação apresentou um R^2 de 87,27% e um R^2_{ajust} de 81,23%.

$$
\begin{aligned}
DLCSF_t = {} & +0,025 - 0,732\,DLCSF_{t-1} - 0,208\,DLCSF_{t-2} + 0,210\,DLCSF_{t-6} - 0,235\,DLCSF_{t-10} \\
& -0,285\,DLCSF_{t-11} - 0,110\,DLCSF_{t-12} - 0,336\,DLTC_t - 0,565\,DLTC_{t-2} - 0,403\,DLTC_{t-4} \\
& +0,725\,DLTC_{t-5} - 0,667\,DLTC_{t-6} + 0,404\,DLTC_{t-7} - 1,506\,DLTT_t + 2,293\,DLTT_{t-7} \\
& +0,882\,DLTT_{t-8} + 0,931\,DLTT_{t-9} + 1,139\,DLTT_{t-10} + 0,112\,DLTJ_{t-5} - 0,155\,DLTJ_{t-7} \\
& -0,127\,DLTJ_{t-10} - 0,168\,DLTJ_{t-11} - 3,263\,DLUCI_{t-2} + 1,144\,DLUCI_{t-7} + 2,097\,DLUCI_{t-9} \\
& +1,272\,DLUCI_{t-10} + 0,676\,DLPFA_t + 0,675\,DLPFA_{t-2} - 0,218\,DLPFA_{t-3} - 1,343\,DLEA-EV-CP_t \\
& -2,126\,DLEA-EV-CP_{t-1} - 2,066\,DLEA-EV-CP_{t-3} - 1,018\,DLEA-EV-CP_{t-4} \\
& -1,315\,DLEA-EV-CP_{t-7} + 1,374\,DLEA-EV-CP_{t-12} - 4,876\,DLECC_{t-9} - 4,430\,DLECC_{t-11} \\
& +4,907\,DLECC_{t-12}
\end{aligned}
$$

No quarto modelo de regressão dinâmica construído, MOD-Geral, que representa a forma "tradicional" de se construir um modelo de regressão, ou seja, sem a análise posterior de componentes principais, utilizou-se a mesma transformação de CSF dos demais modelos. Este modelo contou com todas as 29 variáveis explicativas coletadas e por isso apenas puderam ser consideradas três defasagens de cada uma delas.

O MOD-Geral partiu de um modelo dinâmico geral com 120 parâmetros para 123 observações, o qual após ser estimado forneceu uma equação final, como a apresentada, com apenas 46 parâmetros e com um R^2 de 86,12% e um R^2_{ajust} de 76,95%.

$$
\begin{aligned}
DLCSF_t = {} & -0,728\,DLCSF_{t-1} - 0,364\,DLCSF_{t-2} + 0,503\,DLTC_t + 0,209\,DLARF_t \\
& -0,126\,DLARF_{t-1} - 1,767\,DLDI_t - 0,310\,DLTJ_t - 0,179\,DLTJ_{t-2} + 2,759\,DLIGP-M_{t-1} \\
& -4,281\,DLUCI_{t-1} + 2,874\,DLUCI_{t-3} + 0,436\,DLPFA_{t-2} - 0,858\,DLPFA_{t-3} + 1,341\,DLSM_{t-1} \\
& +0,663\,DLSM_{t-2} - 2,407\,DLEA-EV-CP_{t-2} + 1,865\,DLEA-EV-CP_{t-3} + 9,637\,DLEEM_{t-2} \\
& -11,523\,DLEEM_{t-3} + 17,873\,DLEIT_t + 39,898\,DLEIT_{t-2} - 20,346\,DLEMNM_{t-2} - 12,389\,DLEMNM_{t-3} \\
& -12,608\,DLEM_{t-1} + 18,003\,DLEM_{t-3} - 2,479\,DLEMEC_t - 4,667\,DLEMT_t + 3,996\,DLEMT_{t-1} \\
& -6,754\,DLEMOB_{t-1} - 13,746\,DLEMOB_{t-2} - 5,336\,DLEFC_{t-2} + 13,724\,DLEQPF_{t-1} - 8,870\,DLET_t \\
& -12,537\,DLET_{t-3} - 5,327\,DLEC_{t-1} - 1,873\,DLEPAB_{t-1} - 5,308\,DLEPAB_{t-2} - 16,080\,DLESIUP_{t-1} \\
& -13,978\,DLECC_{t-2} + 9,973\,DLECC_{t-3} + 10,356\,DLECOM_t - 37,738\,DLES_{t-3} + 14,985\,DLEAPA_{t-1} \\
& -8,780\,DLEAPA_{t-2} + 0,267\,DEEGp_{t-1} - 0,254\,DEEGp_{t-2}
\end{aligned}
$$

O desempenho de cada um dos modelos apresentados acima podem ser visualizados na Tabela 4, a qual apresenta um resumo das características de cada um deles.

Modelo	Nº Parâmetros	R^2 ajustado	Erro Padrão	EPAM	U de Theil
MOD-01	29	74,92%	0.064	4,45%	0,423
MOD-02	37	60,60%	0.077	5,97%	0,640
MOD-03	18	81,23%	0.053	3,30%	0,355
MOD-Geral	16	76,95%	0.061	3,98%	0,200

Fonte: PcGive 10 e Aplicativo Excel.

Tabela 4 – Características e desempenho dos modelos

Ao se observar a Tabela 4, pode-se perceber que coerentemente com os pressupostos teóricos,

o modelo que gerou as melhores predições (MOD-03), ou seja, menor EPAM é também o modelo que possui o maior R^2_{ajust} e o menor erro padrão. Apesar de não ser o que possui o maior número de parâmetros, o que justificaria a pressuposição de estar errando menos pelo fato de incluir mais variáveis em sua equação final, portanto, acredita-se que a aplicabilidade da técnica multivariada para estimar a independência entre variáveis antes de se estimarem suas dependências, não ferem os pressupostos básicos que permitem a utilização dos modelos gerados através desta última.

Com base nos resultados encontrados, observa-se que os modelos que apresentaram as menores discrepâncias (MOD-3 e MOD-Geral), os dados mais antigos foram considerados com maior significância no momento da predição do que os valores mais recentes. Tal afirmação baseia-se nos valores do cálculo do U de Theil destes modelos, que se revelaram mais próximos de zero, o que provavelmente esteja ligado ao fato de tratar-se de uma série originalmente auto-regressiva, como pode ser constatado na Figura 3.

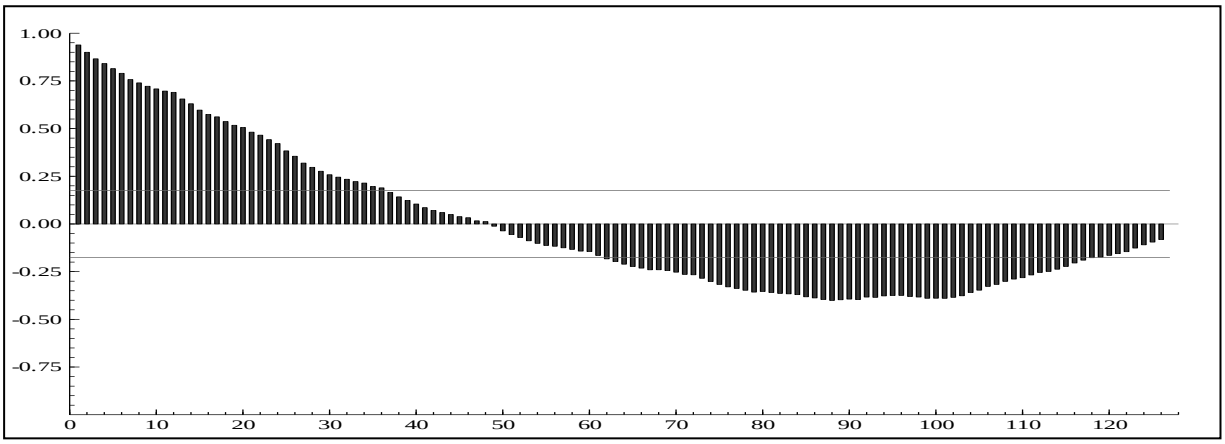

Figura 3 – Correlograma da função de autocorrelação de CSF

As constatações mencionadas são também fundamentadas pelas informações contidas na Tabela 5, que é composta pelos resultados dos testes de especificação dos modelos construídos. Tais resultados, apresentam p valores superiores a 0,05, assumido como nível de significância, para todos os testes realizados, permitindo que se aceitem suas hipóteses nulas de não autocorrelação (teste AR), independência quadrada (ARCH) e normalidade (*Normality*) nos resíduos, além de não heterocedasticidade nas variáveis consideradas (Hetero) e não existência de problemas na especificação da equação (RESET), respectivamente, admitindo, portanto, suas predições como válidas para análise.

Modelo	**AR**	**p valor**	**ARCH**	**p valor**	**Normality**	**p valor**	**Hetero**	**p valor**	**RESET**	**p valor**
MOD-01	1,2452	0,2872	1,0780	0,3852	2,1646	0,3388	0,6561	0,9243	0,1429	0,7063
MOD-02	1,0475	0,4038	1,8825	0,0820	5,9446	0,0512	1,3540	0,1500	2,9593	0,0885
MOD-03	0,8472	0,5523	0,4590	0,8604	4,3974	0,1109	0,0555	1,0000	0,5414	0,4641
MOD-Geral	0,9257	0,4923	0,5010	0,8303	1,2202	0,5433	74,141	0,9136	0,0595	0,8080

Fonte: PcGive 10.

Tabela 5 – Resultados dos testes de especificação

Quanto ao método utilizado para combinação das predições dos modelos construídos, foi

realizada uma média ponderada, como sugerido, por exemplo, por Armstrong (2004) e Evans (2003). Nela se empregou o programa não linear de otimização do aplicativo Excel (Solver), para que se encontrassem os pesos que minimizassem seu EPAM, onde se realizaram os cálculos da soma do produto das predições de cada modelo pelos seus respectivos pesos, dividida pelo somatório destes pesos.

Esta combinação justifica-se, pelo fato de acreditar-se que a utilização da estrutura de ACP só faria sentido se fossem consideradas as diversas dimensionalidades, em conjunto, apontadas por ela como significativas. Estas diversas dimensionalidades são diferentes eixos que representam as direções com máxima variabilidade obtidas do deslocamento e rotação do sistema original de variáveis através de combinações lineares (componentes), e que, segundo Lopes (2001), são justamente eles que fornecem esta descrição mais simples e mais parcimoniosa da estrutura de correlação em estudo.

Aplicando-se então o método de combinação mencionado obteve-se como resultado um maior peso para o MOD-03 (0,687543), seguido do MOD-01, ao qual foi creditado um peso inferior a metade do anterior (0,312457). Quanto ao MOD-02, este, surpreendentemente, recebeu peso zero, mostrando-se dispensável para a construção da resposta combinada. Coincidentemente ou não, este foi o modelo que apresentou o maior erro padrão, o maior EPAM e o que deu mais peso aos dados mais recentes, além de, curiosamente, ser o que inclui o maior número de parâmetros em sua equação final.

3. Conclusões

Os resultados atingidos com a aplicação da metodologia proposta podem ser visualizados na Tabela 6, os quais indicam que é possível se considerar útil à aplicação de uma análise de componentes principais antes de se construir modelos de regressão, quando se trabalha com um grande conjunto de dados. É bem verdade que tal afirmação pode ser feita com segurança apenas para os dados utilizados no estudo e no período em que a metodologia proposta foi aplicada, porém, acredita-se que para o propósito do trabalho, de se avaliar a utilidade da aplicação preliminar de um método que modele a independência entre muitas variáveis antes de se aplicar um método que modele a dependência entre elas, tais resultados constituem-se em evidências que a favorece.

Modelos e Métodos	EPAM	U de Theil
MOD-01	4,45%	0,423
MOD-02	5,97%	0,640
MOD-03	3,30%	0,355
MOD-Geral	3,98%	0,200
Combinação por Média Ponderada	2,93%	0,311

Fonte: Aplicativo Excel.

Tabela 6 – Resultados finais

Como observado na Tabela 6, o resultado atingido pela combinação por média ponderada dos modelos que representam os componentes da ACP, foram mais precisos do que o atingido pelo modelo construído com o "método tradicional", segundo o critério do erro percentual absoluto médio. A utilidade de se trabalhar com a análise preliminar de componentes principais, no caso específico, ficou evidenciada na redução de mais de um ponto percentual nos erros de predição, o que, dependendo da área em que se estiver aplicando tal metodologia, pode significar redução de custos, ganhos de qualidade ou mesmo de produtividade, diferença esta que pode determinar à eficiência ou não de um gestor.

Com relação às vantagens de se aplicar tal metodologia, identificou-se a possibilidade de se incluir um maior número de defasagens nas predições finais, devido ao menor número de

variáveis consideradas por modelo. Por outro lado, isto pode não garantir um maior número de variáveis incluídas nas predições combinadas, quando comparadas com o "método tradicional", como no caso da ilustração realizada.

Mesmo a predição do percentual da segunda devolução de cheques em cada 1.000 compensados não sendo o objeto principal de análise no estudo, algumas constatações podem ser apresentadas, e acredita-se de devem ser feitas até como forma de reforçar a legitimidade da análise realizada. Com relação aos modelos dinâmicos construídos, os resultados apresentados acima em forma de equação, nos permitem identificar as defasagens t-1 e t-2 de CSF, como as mais significativas na tentativa de explicação do comportamento da variação percentual na emissão de cheques sem fundo, na medida em que estiveram presentes nos modelos considerados significativos na análise comparativa. O coeficiente negativo destas defasagens indica que a aversão a cheques deve basear-se em dados de, pelo menos, dois meses atrás.

Por outro lado, podemos ressaltar que, no período analisado, as variáveis Arrecadação Bruta da Receita Federal (ARF); Salário Mínimo Real (SM); Emprego Formal na Indústria de Materiais Elétricos e de Comunicação (EMEC) e Emprego Formal na Administração Pública Direta e Autárquica (EAPA), não persistiram na ACP, e que como o método de combinação adotado deu peso zero para o MOD-02, também não foram incluídas nas predições finais, além destas, o Emprego Formal na Indústria de Materiais de Transporte (EMT), Emprego Formal na Indústria de Produtos Alimentares e Bebidas (EPAB), Emprego Formal nos Serviços Industriais de Utilidade Pública (ESIUP) e o Emprego Formal na Indústria Editorial e Gráfica (EGG).

Já no modelo construído através da "metodologia tradicional", a variável que desapareceu do modelo final foi a referente aos termos de troca (TT), a qual se mostrou significativa no modelo que apresentou o menor EPAM e com um comportamento peculiar. Esta variável apresentou um coeficiente negativo para o tempo "t" e positivo para t-7, t-8, t-9 e t-10, o que pode ser explicado, admitindo-se que a emissão de cheques no Brasil está ligada a compras a prazo, e que tais emissões são de períodos de mais de seis meses e menos de um ano.

Acredita-se que o motivo da influência negativa de TT no tempo atual esteja no fato de que uma deterioração nos termos de troca leva a uma maior devolução de cheques emitidos, pelo fato do mesmo ter sido emitido antes da deterioração. Assim como, um efeito inverso leva a uma redução do percentual de devolução, o que pode ser justificado pela não quebra do planejamento orçamentário do brasileiro, o qual, não esqueçamos, possui uma propensão marginal a consumir superior a um.

4. Agradecimentos

Ao Núcleo de Normalização e Qualimetria (NNQ) do Departamento de Engenharia de Produção e Sistemas (DEPS) que colocou a disposição toda sua estrutura física, incluindo o acesso ao *software* PcGive 10 e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico CNPq – Brasil, pelo apoio concedido através de bolsa de doutorado.

Referências

ARMSTRONG, S, J. (2004). *Principles of forecasting: a handbook for researchers and practitioners.* Massachusetts: Eletronic Services <http://www.wkap.nl>.

CLEMEN, R.T. Combining forecasts: a review and annotated bibliography. **International Journal of Forecasting.** v.5, p.559-583. 1989.

EVANS, M. K. (2003). *Practical Business Forecasting.* Oxford: Blackwell Publishers.

HANKE, J.E.; REITSCH, A.G.; WICHERN, D.W. (2001). *Business Forecasting.* New Jersey: Prentice Hall.

HENDRY, D.F. **Dynamic Econometrics: Advanced Texts in Econometrics.** 1 ed. New York: Oxford University Press Inc., 1995.

HENDRY, D.F., DOORNIK, J.A. **Empirical Econometric Modelling Using: PcGive 10.** vol.1, London: Timberlake Consultants Ltd., 2001.

HENDRY, D.F., RICHARD, J.F. On the formulation of empirical models in dynamic econometrics. **Journal of Econometrics.** v.20, p.3-33. 1982.

LOPES, L.F.D. **Análise de Componentes Principais Aplicada à Confiabilidade de Sistemas Complexos.** Florianópolis, 2001. 121 f. Tese (Doutorado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina (UFSC).

MAHOUD, E. Combining forecasts: some managerial issues. **International Journal of Forecasting.** v.5, p.599-600. 1989.

MAKRIDAKIS, S.G., WHEELWRIGHT, S.C., HYNDMAN, R.J. **Forecasting: Methods And Applications.** 3 ed. New York: John Willey & Sons, 1998.

ZOU, H., YANG, Y. Combining time series models for forecasting. **International Journal of Forecasting.** v.20, p.69-84. 2004.