Escola Nacional de Ciências Estatísticas

Paulo Fernando Mahaz Simões

Modelagem longitudinal de dados de pesquisas de índices de preços ao consumidor, tratamento da não-resposta e índices com heterogeneidade controlada

Rio de Janeiro
2009

Escola Nacional de Ciências Estatísticas

Modelagem longitudinal de dados de pesquisas de índices de preços ao consumidor, tratamento da não-resposta e índices com heterogeneidade controlada

Paulo Fernando Mahaz Simões

Dissertação apresentada ao Programa de Mestrado em Estudos Populacionais e Pesquisas Sociais para obtenção do título de Mestre

Área de Concentração: Metodologia Estatística para Censos, Pesquisas Amostrais e Registros Administrativos

Orientador: Prof. Dr. Djalma Galvão Carneiro Pessoa

Rio de Janeiro
2009

FOLHA DE APROVAÇÃO

À Minha Esposa, Kelly

À Minha Filhinha, Pietra

Ao Meu Sobrinho, Guilherme

AGRADECIMENTOS

É comum, ao conversarmos com pessoas que concluíram trabalhos relativamente longos, como as dissertações de Mestrado ou Teses de Doutorado, que exigem dedicação de tempo ou mesmo momentos de abstração do cotidiano, ouvirmos que a caminhada não teria sido possível sem a contribuição, direta ou indireta, de familiares, amigos e colegas de trabalho. E o meu caso não fugiu à regra. Por isso, gostaria de oferecer um singelo agradecimento a todas essas pessoas.

Quero agradecer a Deus pela força e por tudo.

Quero agradecer aos meus pais, Carlos e Rute, pela educação que me proporcionaram; ao meu irmão Flávio e ao meu sobrinho Guilherme, orgulho de toda a família. Não posso deixar de citar minha sogra, Ruth; Sem sua ajuda para cuidar de minha filhinha, o tempo teria sido muito mais escasso!

Quero oferecer um “muito obrigado” ao IBGE, à Diretoria de Pesquisas, à Wasmália, Zélia e Márcia Quintslr pela oportunidade e incentivo para que ingressasse no Mestrado.

Quero agradecer aos meus colegas e amigos André Costa, Denise Freire, Miguel Bruno, José Fernando, Paulo Medeiros, André Wallace, Guilherme Moreira, Janaina Senna e também a Marcelo Cruz, Gustavo Vitti e Eduardo Wilkinson pelas dicas e sugestões sempre muito pertinentes. Edu, valeu pela releitura do texto!

Quero lembrar dos meus amigos e professores ao longo do curso. Quero também agradecer à Dr.ª Solange Corrêa Onel e à Dr.ª Maria Tereza Serrano Barbosa, integrantes da Banca Examinadora, pelos comentários, críticas e sugestões.

Quero agradecer ao Djalma por toda paciência e por todo o conhecimento transmitido nestes dois anos de orientação.

E, em especial, quero agradecer à minha esposa, Kelly, e à minha filhinha, Pietra, simplesmente, as razões da minha vida.

"Se fosse realizada uma pesquisa com economistas e estatísticos, eles certamente apontariam a falha em não se considerar a mudança de qualidade como uma das mais sérias imperfeições dos índices de preços."
National Bureau of Economic Research (1961, p. 35, tradução nossa)

RESUMO

Neste trabalho, mostra-se como dados de pesquisas de índices de preços ao consumidor (IPC) podem ser analisados segundo um enfoque longitudinal. À luz desta abordagem e no limite dos agregados elementares, são ajustados modelos hierárquicos capazes de captar a estrutura de covariância presente, tais como os modelos de efeitos mistos (MEMs) e os modelos de padrões de covariância (CPM). Estes modelos, além de oferecerem a possibilidade de avaliação de preditores relacionados aos níveis e variações de preços, são instrumentos valiosos para imputação da não-resposta. No caso dos IPCs, a hipótese de não-resposta do tipo completamente aleatória não parece plausível, de modo que os modelos de efeitos mistos, que podem ser utilizados para imputação em situações onde as suposições a respeito das perdas são menos restritivas, sobressaem em relação a outros métodos normalmente empregados. Em complemento ao trabalho, modelos hedônicos com coeficientes aleatórios, incorporando informações sobre "porte" dos estabelecimentos comerciais e "marcas" de produtos, identificados na modelagem longitudinal como fatores relevantes para explicação da variância e dos níveis de preços, são utilizados para cálculo de índices com dispersão controlada. Mostra-se como o índice de Dutot, que não atende ao teste da comensurabilidade, se aproxima do índice de Jevons quando calculado sem o efeito da heterogeneidade, controlada pelos modelos, reforçando a hipótese de que a dispersão dos preços é uma importante fonte das diferenças nos resultados provenientes dessas duas fórmulas, as mais empregadas por institutos oficiais de estatística para cálculos no nível elementar.

Palavras-chave: Modelos Longitudinais, Não-resposta, Heterogeneidade, Modelos Hedônicos, Índice de Jevons, Índice de Dutot.

ABSTRACT

In this work, it is shown how the consumer price index (CPI) data can be analyzed in a longitudinal approach. Restricting the study to the elementary aggregates, hierarchical models that account for the covariance structure of the data, such as mixed-effects models (MEM) and covariance pattern models (CPM), are adjusted. Besides offering the possibility of evaluating predictors related with variation and price levels, these models are valuable tools for the treatment of the non-response. In the CPI context, the hypothesis of missing completely at random non-response is not reasonable. Since mixed-effects models can be used in situations where the missing values assumptions are less restrictive and stand out if compared with other methods usually employed, they are adopted in this work. In addition, hedonic models with random effects are used to calculate controlled dispersion indexes. It is shown that the Dutot index, which fails the commensurability test, is close to the Jevons index when calculated without heterogeneity effects controlled by the models. The results reinforce the hypothesis that the price dispersion is an important source of differences between these two formulae, the most commonly adopted by national statistical offices for the calculus of CPIs at the elementary level.

Key-words: Longitudinal Models, Non-response, Heterogeneity, Hedonic Models, Jevons Index, Dutot Index.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ACV	*Autocorrelation Function*
AIC	*Akaike Information Criterion*
AR	*Autoregressive*
ARMA	*Autoregressive Moving Average Model*
CEMPRE	*Cadastro Central de Empresas*
CNAE	*Classificação Nacional da Atividade Econômica*
COICOP	*Classification of Item Consumption by Purpose*
CPI	*Consumer Price Index*
CPM	*Covariance Pattern Models*
EQM	*Erro Quadrático Médio*
FAC	*Função de Auto-correlação*
IBGE	*Fundação Instituto Brasileiro de Geografia e Estatística*
ICV	*Índice de Custo de Vida*
ILO	*International Labour Organization*
IPC	*Índice de Preços ao Consumidor*
IPCA	*Índice Nacional de Preços ao Consumidor Amplo*
ISIC	*International Standard Industrial Classification of All Economic Activies*
lm	*Linear Models, "Biblioteca" do*
lme	*Linear Mixed-effects Models, "Biblioteca" do R*
MA	*Moving Average*
MAR	*Missing at Random*
MCAR	*Missing Completely at Random*
MEM	*Modelos de Efeitos Mistos*
MNAR	*Missing Not at Random*
MQO	*Mínimos Quadrados Ordinários*

MV	*Máxima Verossimilhança*
PLC	*Pesquisa de Locais de Compra*
POF	*Pesquisa de Orçamentos Familiares*
PROD	*Produto*
R	*Software Estatístico*
R$	*Reais*
SNIPC	*Sistema Nacional de Índices de Preços ao Consumidor*
var	*Variância*

SUMÁRIO

Capítulo 1 – Introdução

Dentre as técnicas estatísticas de inferência, as de modelagem apresentam-se como valiosas para o estudo da relação entre variáveis populacionais. Neste trabalho, onde dados de pesquisas de índices de preços ao consumidor (IPC) são modelados segundo um enfoque longitudinal, os resultados obtidos são usados para imputação da não-resposta e para o cálculo de índices com heterogeneidade controlada por intermédio de modelos hedônicos[1].

Considerando-se que estratégias distintas podem ser planejadas para a coleta de dados, duas abordagens se destacam na concepção dos modelos: a transversal, ou *cross-sectional*, quando a amostra é coletada num instante único do tempo; e a longitudinal, quando os dados são observados em mais de um momento para as mesmas unidades amostrais. Adicionalmente, os modelos estatísticos, escritos como combinações de variáveis explicativas, podem incorporar tanto efeitos fixos, referentes a toda população, como aleatórios, relacionados a elementos experimentais extraídos aleatoriamente do universo em estudo. Dependendo do contexto, das características dos dados, objetivos e estratégias de análise, uma ou outra abordagem será a mais apropriada.

Hedeker e Gibbons (2006) citam, entretanto, algumas vantagens dos modelos longitudinais frente aos transversais: primeiro, as medidas repetidas de cada unidade amostral servem para seu próprio controle. Como, em geral, a variabilidade intraindividual é menor que a variabilidade entre indivíduos, os testes estatísticos são mais robustos; em segundo lugar, o pesquisador pode separar efeitos temporais ao longo de um período dos efeitos de *coorte* e, dessa forma, os estudos longitudinais são capazes de propiciar informações sobre

[1] Modelos hedônicos são modelos estatísticos onde os preços dos produtos são regredidos em suas próprias características. Por exemplo, preços de TVs em características como *tipo* e *tamanho da tela*, *qualidade de som*, etc. Os coeficientes das variáveis são percebidos como estimativas dos preços marginais de diferentes características dos produtos. (ILO et al., 2004, p.374).

mudanças individuais, enquanto que as amostras obtidas em momento único carecem dessa prerrogativa.

Os modelos longitudinais são uma vertente dos *modelos hierárquicos* ou *multinível*. Estes modelos são aplicáveis a dados que apresentam determinadas estruturas hierárquicas, tais como estudantes em classes, classes em escolas, operários em indústrias, ou mesmo medidas longitudinais como, por exemplo, as do conjunto de preços mensais de um determinado produto comercializado numa cidade ao longo dos meses. Estas arquiteturas retratam as fontes de variância presentes. Segundo Snijders e Bosker (1999), a abordagem multinível é uma metodologia voltada para a análise de estruturas com padrões complexos de variabilidade onde o foco está, exatamente, nestas fontes de variância. Na especificação de um modelo hierárquico é possível descrever o valor assumido pela variável resposta como uma combinação de efeitos dos grupamentos nos quais as unidades portadoras de informação se inserem.

Outra diferença entre as abordagens transversal e longitudinal reside no fato de que, no caso dos modelos transversais, em geral, a variável dependente y pode ser modelada a partir de uma estrutura de médias $X\beta$, com a suposição de erros normais independentes e com variância constante, $\varepsilon_i \sim N(0,\sigma^2)$, sem a necessidade de considerações a respeito de alguma eventual estrutura de covariância presente. Na modelagem longitudinal, onde são observados valores repetidos no tempo da mesma unidade amostral, estas suposições não são adequadas, já que é plausível supor alguma correlação nas medidas. Os modelos hierárquicos longitudinais, como os que aqui se pretende aplicar, permitindo a incorporação de coeficientes aleatórios em suas estruturas como, por exemplo, interceptos (b_{0i}) e coeficientes angulares (b_{1i}), oferecem, ainda, a possibilidade de avaliação de trajetórias individuais de crescimento.

No contexto dos índices de preços ao consumidor, isto significa a possibilidade de avaliação das curvas de evolução dos preços de cada produto integrante da amostra.

Uma outra classe de modelos, denominada *modelos de padrões de covariância* (CPM)[2], ainda dentro do enfoque longitudinal, considera que a estrutura de variância e covariância dos erros pode ser diretamente modelada a partir de algumas suposições a seu respeito, a saber, autocorrelação de ordem n (AR), média móvel de ordem n (MA) ou mesmo estrutura de correlação de Toeplitz, dentre outras.

Aliás, a aplicação destas estruturas aos erros dos modelos hierárquicos com efeitos aleatórios são um complemento interessante ao processo de estimação. Neste trabalho, estes modelos são estendidos com o intuito de englobá-las e os resultados são comparados com os dos *modelos de padrões de covariância*.

No enfoque estocástico de pesquisas de IPC os modelos longitudinais apresentam-se como notáveis instrumentos de análise, haja vista os fortes indícios de correlação nas variáveis e o reconhecido caráter longitudinal dos dados. Recorrer a tais modelos pode ser útil tanto para avaliação e conhecimento do fenômeno em estudo quanto para predição. Além disso, as estimativas de tendência podem ser usadas para compreensão da heterogeneidade de produtos e locais e investigação dos fatores determinantes dos diferentes níveis e mudanças de preços.

As pesquisas de índices de preços, realizadas na maior parte dos países por institutos oficiais de estatística têm, como propósito mais geral, a obtenção de uma medida síntese do movimento dos preços dos bens consumidos pelas respectivas populações-objetivo em determinado período. Sendo pesquisas complexas, acabam por demandar, a cada dia, atenção especial por parte de analistas e pesquisadores, podendo chegar a ter diversas etapas, como a

[2] Do inglês *Covariance Pattern Models*.

realização das pesquisas de orçamentos familiares (POFs)[3], que fornecem estimativas da importância de cada bem no consumo das famílias; as de estabelecimentos comerciais (PLCs)[4] , para cadastro de produtos comercializados e, ainda, levantamentos mensais, cuja finalidade é a coleta de preços.

Para a divulgação de um indicador mensal são indispensáveis, além da mensuração dos preços em estabelecimentos comerciais, a crítica e a imputação de dados, o cálculo e a análise dos resultados, etapas nas quais, com freqüência, o pesquisador depara-se com problemas que exigem soluções rápidas e adequadas. Ou seja, as pesquisas de IPCs constituem um enlace de distintos processos, sujeitos inclusive à aleatoriedade, onde a utilização de técnicas estatísticas, das quais a modelagem longitudinal é uma delas, pode ser de grande serventia.

O processo de cálculo de um IPC se desenvolve, basicamente, em dois estágios. No primeiro, a partir de uma amostra de preços provenientes de estabelecimentos comerciais e prestadores de serviços, são calculados índices para os agregados elementares e, em seguida, num segundo estágio, os índices obtidos para esses agregados elementares são então combinados para obtenção de índices em estágios mais elevados de agregação. Estes estágios mais elevados, diferentemente do que acontece no nível elementar, são os que, de acordo com o padrão internacional, recebem ponderações oriundas das pesquisas de orçamentos familiares.

Os agregados elementares, onde se propõe a aplicação dos modelos aqui estudados, consistem em estratos homogêneos de produtos normalmente bem definidos e agrupados de acordo com algum sistema de classificação. Pela forma como são estabelecidos e pela arquitetura da estrutura de classificação, acabam por serem considerados como estratos ou,

[3] Ver Sistema (2005) para informações sobre a POF 2002/2003, realizada pelo IBGE.
[4] Em Sistema (2007) há informações sobre a PLC realizada pelo IBGE.

pelo menos, como variáveis de grande importância no momento de dimensionamento das amostras dos IPCs.

A propósito da adoção da abordagem estocástica no trato dos agregados elementares, Carmo (2004) comenta que, apesar da Teoria Econômica do Consumidor ser a mais relevante para as etapas mais agregadas de cálculo em IPC, o enfoque estocástico, no qual o IPC é tratado como medida de tendência central da distribuição dos preços relativos, ao lado do axiomático, tem sido a referência principal.

Entretanto, é importante frisar, este trabalho não busca investigar qualquer contradição que por ventura possa existir entre as teorias estocástica, econômica e axiomática. Pelo contrário, todo o esforço converge para a busca de complementaridade entre as três concepções.

A modelagem dos preços, vinculada ao enfoque estocástico, pode ser relevante sob alguns aspectos: num primeiro momento, a análise estatística de variáveis socioeconômicas pode auxiliar a compreensão do movimento dos preços dos produtos e serviços, permitindo identificar a influência de determinados preditores nos níveis de preços praticados por estabelecimentos comerciais; num segundo espectro de investigação, valores preditos com base nos modelos estimados podem ser empregados para imputação da não-resposta. A esse respeito, deve-se destacar que em alguns levantamentos estatísticos a não-resposta pode ser ignorada sem que conclusões ou resultados sejam comprometidos significativamente. Contudo, em certas pesquisas, nas quais estão incluídas as de índices de preços, a ocorrência de não-resposta pode constituir problema vultoso, exigindo cuidados especiais não somente em seu trato mas, também, na suposição a respeito do mecanismo gerador das perdas. Este estudo salienta esta precaução. Reside, aqui, mais uma vantagem dos modelos longitudinais que, estimados com base em verossimilhança, são robustos a situações onde outras técnicas utilizadas para imputação de não-resposta gerariam, por exemplo, estimativas viciadas.

Ainda no âmbito dos índices elementares, três fórmulas de cálculo são empregadas internacionalmente por institutos oficiais de estatística: a de Dutot, que consiste numa razão de preços médios entre dois períodos; a de Jevons, média geométrica das razões de preços; e a de Carli[5], que equivale a uma média aritmética de relativos, mas que está em desuso por ser viciada e não atender ao teste da reversibilidade temporal[6].

Com relação às fórmulas de Jevons e de Dutot, segundo Silver e Heravi (2007)[7], apesar da escolha entre uma ou outra normalmente ser baseada nas teorias econômica e axiomática, a dispersão dos preços tem papel determinante na diferença dos resultados gerados pelas duas formulações. A hipótese por eles levantada é a de que se a variabilidade dos preços decorrente da heterogeneidade de fatores atrelados tanto aos produtos em si quanto aos estabelecimentos comerciais puder ser controlada, o índice de Dutot, calculado por modelos hedônicos, será mais próximo do índice de Jevons do que o tradicionalmente calculado. Logo, aparece aqui a terceira importância da modelagem proposta: preditores determinantes da variância dos preços podem ser empregados para o cálculo de índices de Jevons e de Dutot com heterogeneidade controlada, mais próximos da "variação pura de preços". É o que se propõe nesta dissertação como complemento a estimação dos modelos longitudinais.

Na busca pelos objetivos gerais aqui descritos e para embasar o desenvolvimento metodológico, são utilizados dados referentes ao subitem "arroz", pesquisados mensalmente nos levantamentos do Sistema Nacional de Índices de Preços ao Consumidor (SNIPC)[8], do IBGE. Pretende-se, contudo, que as avaliações relacionadas à modelagem longitudinal

[5] Existem outras opções, menos usadas, como a adotada em seis países, dentre os quais o Brasil, que consiste num cálculo de razão de médias num primeiro estágio e aplicação de uma média geométrica dessas razões num segundo momento (Silver e Heravi, 2007).
[6] No segundo capítulo, estes índices, bem como algumas de suas propriedades, são apresentados de forma mais detalhada.
[7] O tema foi apresentado pelos autores em versão preliminar no sétimo encontro do International Working Group on Price Indices (SILVER e HERAVI, 2003).

possam ser estendidas a outros itens de pesquisas de índices de preços ao consumidor. Enfatiza-se, ainda, que não há, necessariamente, qualquer vinculação entre dados oficiais do SNIPC/IBGE e o estudo acadêmico aqui proposto, onde a apresentação das possibilidades oferecidas às pesquisas de índices de preços ao consumidor pela abordagem longitudinal e as aplicações dela decorrentes constituem o interesse primordial, mais relevantes que resultados particulares em si obtidos.

1.1 – Objetivos Específicos

A modelagem de dados de pesquisas de índices de preços ao consumidor com o auxílio de modelos longitudinais e das perspectivas por ela oferecidas, quais sejam, imputação e explicação do comportamento de preços em função de preditores selecionados, a partir da consideração de que existe uma estrutura de covariância presente nos dados e de que um mecanismo gerador de perdas pode agir no processo mensal de formação do painel de preços, constitui o eixo central deste trabalho. Em complemento ao estudo, são avaliadas as diferenças entre os índices de Jevons e de Dutot calculados de forma tradicional e as diferenças entre esses mesmos indicadores quando o índice de Dutot é calculado a partir de modelos hierárquicos hedônicos, sem o efeito de fatores determinantes de variabilidade, identificados na modelagem longitudinal.

[8] Sistema (2007).

1.2 – Justificativas

"O pesquisador deve sempre estar ciente de que os resultados da pesquisa são apenas tão bons quanto a qualidade dos dados" (GUJARATI, 2000, p.15). Esta frase ilustra a importância do cuidado que se deve ter com a base de dados em ciências sociais. Indo um pouco além, abstraindo-se das pesquisas de índices de preços e transportando a mensagem para os dados de entrada em qualquer estudo, seja ele social, econômico ou relacionado a qualquer outro campo do conhecimento, não é difícil perceber como a credibilidade dos trabalhos deve depender da qualidade da base de informações, dos dados utilizados e dos procedimentos analíticos implementados.

Os índices de preços ao consumidor são estatísticas cujos usos vão muito além da sua função aparentemente simples de medidor da inflação ou de auxílio ao monitoramento de políticas macroeconômicas. São indicadores também usados para outros fins, entre eles, por exemplo, o de servir como parâmetros para reajustes contratuais e salariais. Nas áreas econômica e social, muitas pesquisas não podem prescindir de seus resultados, já que os utilizam como deflatores em seus processos de análise e obtenção de valores reais.

Estas breves palavras retratam, espera-se, a importância dos índices de preços ao consumidor não só como estatísticas com um fim imediato e bem determinado mas, também, como insumo em outros processos, o que demonstra a necessidade de aprimoramentos constantes da metodologia envolvida em seus cálculos. Tão importante quanto seus usos é a complexidade, já citada, das pesquisas de IPCs, nas quais um problema recorrente é, exatamente, a não-resposta.

Vinculada a estes comentários, a proposta apresentada nesta dissertação, a de modelagem de preços para fins de imputação e para avaliação de variáveis a eles correlacionadas por modelos que prezam pelo caráter longitudinal dos dados e são robustos à

ocorrência de não-resposta, pode constituir-se num instrumento adicional a ser posto à disposição de analistas para a realização de seus trabalhos.

1.3 – Organização do texto

Após este capítulo introdutório, o segundo capítulo aborda aspectos teóricos relacionados aos índices de preços ao consumidor. No terceiro, comenta-se a ocorrência de não-resposta em pesquisas de índices de preços e as consequencias do uso de algumas técnicas de imputação diante dos diferentes tipos de não-resposta. Uma análise descritiva dos dados utilizados no desenvolvimento metodológico é feita no quarto capítulo. No quinto, realiza-se a modelagem longitudinal. São discutidos os modelos hierárquicos e os modelos de padrões de covariância. No capítulo seis, os resultados de imputação obtidos via modelagem longitudinal são comparados com os gerados por outros três métodos de imputação adotados em pesquisas de IPC e que são: 1) imputação pelo preço médio dos produtos semelhantes comercializados no mês; 2) imputação pelo último preço observado; e 3) imputação pela variação média dos preços dos produtos observados. A extensão da modelagem para uma amostra mais heterogênea de produtos é feita no capítulo número sete, juntamente com um estudo sobre as diferenças entre os índices de Jevons e de Dutot com dispersão controlada a partir de modelos hedônicos. Por fim, apresentam-se comentários sobre o trabalho desenvolvido, suas limitações e possibilidades de extensão.

Capítulo 2 – Índices de Preços ao Consumidor

Segundo o disposto no manual de índices de preços ILO et al. (2004), os índices de preços ao consumidor são indicadores que medem a taxa de mudança dos preços de bens e serviços de consumo durante dois instantes. Em outras palavras, os IPCs referem-se à variação dos preços de bens e serviços consumidos por um grupo de indivíduos, denominado "população-objetivo", durante um período de tempo estabelecido.

Os índices de preços ao consumidor tiveram origem no século XVIII. Os primeiros índices adotados foram os de Laspeyres e Paasche, em torno de 1870. Entretanto, uma discussão sobre IPCs não se completa sem uma devida alusão aos índice de custo de vida, ICVs. E, vale ressaltar, apesar de no campo teórico índice de preços ao consumidor e índice de custo de vida caminharem lado a lado, na realidade, são dois conceitos distintos. Entretanto, ao longo dos anos, muitos foram os que confundiram as duas nomenclaturas, utilizando-as com o mesmo significado, ainda que suas diferenças fossem cristalinas. É o que se depreende da citação de Melo (1976): "A sinonímia atribuída principalmente a Custo de Vida e Índice de Preços ao Consumidor está longe de ser verdadeira."

O ICV, conceito elaborado pelo economista russo Konüs em 1924[9], refere-se à comparação de despesas monetárias às quais os consumidores estariam submetidos caso desejassem manter suas preferências entre dois momentos, onde estas preferências seriam dadas por curvas de indiferença[10].

Viabilizar o cálculo de um índice de custo de vida seria extremamente complexo já que, dentre outros elementos, seria necessário o conhecimento das preferências individuais em dois instantes, o que hoje seria inviável, não só pela própria dificuldade de identificação das

[9] Konüs (1939)
[10] Um texto de referência sobre curvas de indiferença pode ser encontrado em Varian (1993).

preferências individuais em si, quanto pela dificuldade que haveria em se pesquisar preços de tão ampla gama de produtos ou serviços. Além do mais, seria preciso garantir a estabilidade das preferências no período analisado.

Já os índices de preços ao consumidor, normalmente considerados como aproximações dos ICVs, podem se pautar, por exemplo, na suposição de que o nível de satisfação dos consumidores se mantêm constante ao longo de um determinado período de tempo. É esta, aliás, a suposição em que se baseia a fórmula de Laspeyres, muito utilizada para cálculo de IPC em diversos países. Enquanto no ICV a pesquisa de todas as especificações de produtos e serviços seria indispensável, no IPC uma combinação de preços de alguns itens de despesas pode representar seu resultado mensal. Melo (1976) apresenta um estudo formal das diferenças entre IPC e ICV.

A estreita relação entre os termos IPC e inflação é outro ponto que merece especial atenção. Segundo Sáinz e Manuelito (2006), deve-se observar que os bens e serviços existentes numa economia são em número muito maior do que os usualmente pesquisados em índices de preços ao consumidor, uma vez que estes se propõem a serem representativos apenas do consumo privado. Lembra-se, portanto, que o conceito de inflação, a rigor, pode circunscrever um leque mais completo de bens, incluindo alguns de consumo duráveis não pesquisados em IPC, bens de capital, ativos financeiros e bens imóveis.

Entretanto, não obstante o fato dos produtos em IPC estarem limitados a uma parcela restrita do conjunto global de bens e serviços produzidos e disponibilizados no mercado, os índices de preços são as estatísticas mais usadas para retrato da inflação. Principalmente em países que observaram extensos períodos de alta de preços, como o Brasil nas últimas décadas do século passado, os indivíduos se acostumaram a perceber os IPCs como dentre os mais importantes indicadores econômicos. O comentário seguinte corrobora o texto acima:

> [...] Os índices de Preços ao Consumidor são, provavelmente, as estatísticas econômicas divulgadas com maior freqüência e destaque no Brasil. Mesmo os indicadores de desemprego e de produto não merecem a mesma atenção na mídia. Isto reflete a relevância desse indicador para a vida de pessoas e instituições, uma vez que o monitoramento da inflação é fundamental para a política monetária e muitos contratos são corrigidos monetariamente por índices de preços. (CARMO, 2004).

Sendo estatísticas de amplo uso, os IPCs são também tratados como das mais relevantes a serem disponibilizadas para o planejamento de políticas macroeconômicas, haja vista o caso brasileiro do Índice Nacional de Preços ao Consumidor Amplo (IPCA), produzido mensalmente pelo IBGE e adotado pelo Banco Central do Brasil como balizador do regime de metas de inflação[11]. Além disso, o IPCA é utilizado oficialmente para reajustar o salário mínimo.

Dentre os principais usos dos índices destacam-se, além do emprego destes como medidas de inflação e parâmetros para reajustes contratuais, outras importantes aplicações, tais como instrumentos para deflacionamentos na contabilidade nacional e em pesquisas como as das áreas de comércio, serviços e de orçamentos familiares, realizadas pelo IBGE. No campo social, sobressaem estudos relacionados a comparações de preços em níveis regional e internacional. Os índices são também utilizados, em muitos países, para correções de aposentadorias e benefícios sociais. No Brasil, uma série de tributos municipais, estaduais e federais são corrigidos com base nos resultados destes indicadores.

[11] Sistema (2005).

2.1 – Amostragem e Pesquisas de Preços

Os preços, nas pesquisas de IPC, são geralmente coletados com periodicidade mensal[12]. São obtidos com base em amostras de estabelecimentos comerciais construídas por processos que variam muito de um país para outro. Por exemplo, na Suécia e nos Estados Unidos a maioria dos subitens é pesquisada segundo critérios de seleção probabilística tanto para locais quanto para produtos[13]. Já em países como Austrália e Canadá a seleção da amostra é intencional, com a seleção dos produtos obedecendo a critérios particulares.

No caso brasileiro do SNIPC, as revisões amostrais já permitiram a seleção de locais com base em critérios probabilísticos. Entretanto, a reposição amostral das perdas de estabelecimentos comerciais que vem sendo feita ao longo dos anos tem obedecido a critérios não probabilísticos, prática esta que não difere da observada na maioria dos países, principalmente nos das Américas do Sul e Central, Ásia, África e em alguns países europeus. Um dos argumentos frequentemente mencionados para justificar processos não probabilísticos é a falta de cadastros de locais e produtos[14].

Abordando-se mais especificamente o SNIPC, fonte de dados deste estudo, a coleta de preços é realizada todos os meses nas regiões metropolitanas do Rio de Janeiro, Porto Alegre, Belo Horizonte, Recife, São Paulo, Belém, Fortaleza, Salvador e Curitiba, além do município de Goiânia e do Distrito Federal. Os locais que compõem a amostra englobam

[12] Há subitens e produtos que não são pesquisados todos os meses. Dependendo do dimensionamento amostral e dos critérios utilizados, a coleta pode ser bimensal, trimestral, semestral e até mesmo ter sua periodicidade variando em função da área geográfica da pesquisa.
[13] Para informações sobre o IPC sueco: "The Swedish Consumer Price Index" (2001).
[14] ILO et. al (2004, p.72) cita as principais justificativas apresentadas por aqueles que defendem seleções não-probabilísticas: falta de cadastros, entendimento de que vícios são irrelevantes, limitações orçamentárias, etc. O mesmo texto contrapõe todos esses argumentos.

estabelecimentos varejistas diversos, como supermercados, farmácias e prestadores de serviços, dentre outros[15].

2.2 – Estrutura de Pesos e Sistema de Classificação

Os IPCs resultam, via de regra, da agregação de dois subsistemas de informações, um de preços e outro de pesos. Os preços são provenientes de levantamentos mensais realizados em estabelecimentos comerciais e prestadores de serviços e, os pesos, originados a partir de pesquisas de orçamentos familiares, nas quais são calculadas as participações relativas de cada bem ou serviço na despesa total das famílias. O resultado final de um IPC representa, portanto, uma média ponderada de razões de preços[16], onde as importâncias relativas dos produtos atuam como pesos.

A estrutura de pesos, também denominada estrutura de ponderações, consiste numa organização hierarquizada de bens e serviços de consumo, onde cada bem de consumo ou item de despesa aparece com sua respectiva participação na despesa média familiar. Essas participações são calculadas, normalmente, com a realização de pesquisas de orçamentos familiares.

Em geral, a arquitetura de uma estrutura de ponderadores é desenhada com base em algum critério de classificação dos bens e serviços que compõem a cesta de consumo da população-objetivo. Internacionalmente, a classificação amplamente empregada por institutos oficiais de estatística para determinação da composição dos grupos e subgrupos das estruturas

[15] Para o trabalho aqui realizado, considerou-se dados de preços de arroz pesquisados em supermercados. A suposição subjacente é que a amostra é representativa da população de supermercados na região metropolitana do Rio de Janeiro.
[16] A seção 2.2 apresenta mais detalhes sobre as fórmulas de cálculo mais utilizadas em IPCs.

de pesos é a "Classificação de Consumo Individual por Finalidades" (COICOP)[17]. Neste sistema, os produtos são agrupados de acordo com o critério da finalidade (ou propósito) de seu uso.

Apesar da COICOP ser o padrão internacional, não é a única possibilidade de organização de uma estrutura de bens e serviços. Outros critérios, como o do "tipo de produto", onde os elementos são arranjados segundo suas características físicas ou natureza, como é o caso da "*Central Product Classification*", ou ainda, o critério da classificação segundo a atividade econômica que deu origem ao produto, adotado pela *International Standard Industrial Classification of All Economic Activies* (ISIC)[18], são outras alternativas. No SNIPC, a classificação adotada é semelhante à COICOP[19].

As estruturas de classificação, divididas em níveis hierárquicos, têm ainda a esperada característica de possuírem pesos maiores para níveis mais agregados. O cálculo dos índices se faz, ordinariamente, para os estágios hierárquicos principais do sistema de classificação, que pode contar com grupos, subgrupos, itens, subitens, etc. A título ilustrativo, uma estrutura organizada a partir de produtos como marcas específicas de cereais, tubérculos, refrigerantes, e outros itens de consumo, poderia ter o aspecto do Quadro 1.

Quando do cálculo mensal de um indicador, os preços ou razões de preços (relativos) pesquisados são combinados com os pesos, gerando, dessa forma, o resultado final, ou global, do IPC. Para apreciação, apresenta-se, anexa, a estrutura completa de pesos do IPCA , obtida por intermédio da Pesquisa de Orçamentos Familiares (2002-2003), do IBGE, para a região metropolitana do Rio de Janeiro[20] (ver anexo 1).

[17] Sigla oriunda do inglês, "Classification of Item Consumption by Purpose". Pode ser consultada no site da Organização da Nações Unidas, *http://unstats.un.org/unsd/cr/registry/regct.asp?Lg=1*

[18] United Nations Publication (2009).
[19] United Nations, Methods and Classifications (2009).

Quadro 1 – Ilustração de uma estrutura de pesos hipotética para IPC

Nível Hieráquico	***Pesos provenientes de uma POF***
Índice geral	100%
Grupo: ***alimentação***, habitação, educação, etc.	25%
Subgrupo: ***alimentação no domicílio***, alimentação fora	20%
Item: ***cereais***, frutas, artigos de limpeza, etc	5%
Subitem: ***arroz***, feijão	1%
Produtos: ***marca A***., ***marca B***,..., (nível elementar)	Sem pesos explícitos da POF

Fonte: Construção do autor com valores hipotéticos

Uma observação atenta da última linha do Quadro 1 revela, porém, que os produtos especificados correspondentes ao nível elementar não têm, via de regra, pesos originários das pesquisas de orçamentos familiares. Não é viável, hoje, em muitos países, que uma pesquisa de orçamentos familiares forneça pesos para tão vasta quantidade de marcas e especificações.

Entretanto, a análise nesse nível, o mais desagregado da estrutura e que resulta na variação de preços do subitem (a combinação das diferentes cotações de tipos e marcas de feijão resulta no índice do subitem "Feijão") é uma das mais importantes para o cálculo de um índice de preços ao consumidor, pois os produtos ali contidos são os reais portadores da informação a ser coletada, o preço, e todas as contas realizadas nas etapas acima deste nível são influenciadas pelas decisões e suposições nele determinadas. A esse respeito, várias são as observações na literatura: "A qualidade do IPC depende fortemente da qualidade dos índices elementares, base para sua construção". (ILO et al. 2004, p.355, tradução nossa). Ou ainda:

[20] Sistema (2005).

> [...] quando analisamos em detalhe a metodologia de um IPC há variantes desses modelos que se aplicadas poderiam levar a diferenças nos resultados obtidos. Quanto à estrutura de ponderações há alternativa de determiná-la segundo um critério plutocrático, em que a cada consumidor seria atribuído um peso proporcional à participação de seus gastos no conjunto de consumidores, ou um critério democrático, em que todos os consumidores teriam implicitamente o mesmo peso. *Mais significativo que isto para explicar diferenças nos resultados é a adoção de fórmulas diferentes para o cálculo de índices elementares, ou seja, o relativo de preços de cada especificação elementar de produto ou serviço*. (CARMO, 2004). [*grifo nosso*]

Logo, é possível perceber como o entendimento dos conceitos e hipóteses subjacentes ao cálculo de um IPC no nível elementar, e não somente nos níveis mais agregados, é fundamental para a compreensão do comportamento dos preços e, principalmente, para realização de inferências a partir das bases de dados.

Nas seções seguintes, abordam-se, de forma sumária, os principais conceitos atrelados às diferentes fórmulas de cálculo usualmente adotadas no nível elementar de análise, bem como as teorias econômica, estocástica e axiomática, com ênfase, porém, nas duas últimas.

O nível elementar de cálculo do IPC recebe especial conotação por ser nele que se propõe a aplicação dos modelos longitudinais e a avaliação das diferenças que surgem nos resultados das fórmulas de Jevons e de Dutot sob duas situações: uma primeira quando o índice de Dutot é calculado segundo os métodos tradicionalmente adotados pelos institutos oficiais de estatística e, outra, quando calculado por modelos hedônicos, com controle da heterogeneidade presente nos dados. Esta metodologia aproxima mais sua estimativa da obtida com o índice geométrico e reforça a tese de que parcela considerável das diferenças entre as duas fórmulas deve-se à dispersão dos preços.

2.3 – Fórmulas de Cálculo e as Abordagens Econômica, Axiomática e Estocástica para o Nível Elementar de Análise

De acordo com Diewert (1995), três são as teorias que pautam a construção dos IPCs: a abordagem econômica, baseada no conceito teórico de custo de vida[21] e atrelada à teoria do consumidor; a axiomática, estabelecida por Fisher (1922) e que propõe a avaliação da fórmula do indicador de variação de preços em função do atendimento a determinados axiomas e propriedades e, por fim, a estocástica, que percebe o IPC como uma medida de tendência central da distribuição das razões de preços (preços relativos)[22].

Contudo, como apontado por Carmo (2004), nenhuma destas abordagens é capaz de fornecer soluções incontestáveis para estimação de índices de preços e, em consequência, aspectos dos três enfoques acabam por serem verificados em várias etapas e em diversos sistemas de cálculo de IPC. No estabelecimento das fórmulas para cálculo acima do nível elementar, há uma prevalência da abordagem econômica em relação às duas outras. A Teoria Econômica do Consumidor tem pautado as etapas mais agregadas de cálculo, enquanto que no nível elementar de análise, foco desta dissertação, as abordagens estocástica e axiomática têm sido mais observadas.

Não existe, portanto, uma fórmula única para cálculo dos índices. Dependendo da abordagem preferida, haverá uma ou outra fórmula mais adequada a ser empregada. Desse modo, as fórmulas podem diferir tanto em decorrência do nível de agregação no qual são aplicadas quanto em função da abordagem teórica subjacente e, frequentemente, estão atreladas a suposições e objetivos pré-estabelecidos. Salienta-se, porém, que, em termos de efeitos nos indicadores, as distintas fórmulas implicam em diferentes resultados.

[21] Para aprofundamento, ver Diewert e Nakamura (1993);

A esse respeito, Silver e Heravi (2007), identificaram, entretanto, que, apesar das teorias econômica e axiomática terem preponderado ao longo dos anos para escolha das fórmulas, a dispersão dos preços relativos é uma importante fonte de divergência dos indicadores de inflação quando estes são calculados por diferentes métodos, teoria esta que, uma vez aceita, releva a abordagem estocástica.

Nas próximas seções, descreve-se a construção dos índices elementares e as fórmulas de Dutot, Jevons e Carli[23]. Maior ênfase, porém, é dada aos índices de Dutot e Jevons, os mais adotados por institutos oficiais de estatística, e às concepções estocástica e axiomática, como mencionado anteriormente. Sobre o índice de Carli, apresentam-se breves comentários e, a respeito da abordagem econômica, um texto de referência é Diewert (1993).

2.3.1 – O Índice de Carli

O índice de Carli[24] entre os tempos t e t-1 pode ser obtido com o cálculo da média de relativos de preços entre os dois períodos. Sua fórmula é:

$$I_C(p^{t-1}, p^t) = \frac{1}{K}\sum_{1}^{K}\frac{p_k^t}{p_k^{t-1}} \qquad (2.1)$$

onde:

$I_C\left(p^{t-1}, p^t\right)$ é o índice de Carli entre os períodos t-1 e t

p_k^t é o preço do bem k no mês t

[22] Segundo Diewert e Nakamura (1993, p. 179), antes da econômica e axiomática, a abordagem estocástica prevaleceu entre 1875 e 1925.

[23] Outras formulações, como a Média Harmônica de Relativos, $I_H(p^0,p^1)$, que assim como índice de Carli não atende ao teste da reversibilidade temporal e a Média Geométrica dos Índices de Carli e da Média Harmônica de Relativos, $I_{CSWD}(p^0,p^1)$, não apresentadas aqui, podem ser encontradas em ILO (2004, p.360-361).

[24] Proposto por Gian Rinaldo Carli em 1764.

p_k^{t-1} é o preço do bem k no mês t-1

K é o número de bem específicos no agregado elementar

De acordo com ILO et al. (2004), o índice de Carli, que já foi amplamente utilizado, vem deixando de ser empregado em decorrência de não atender ao teste da reversibilidade temporal[25] e por ser viciado, levando a resultados superestimados quando as variações de preços são positivas e a resultados subestimados quando são negativas[26]. Neste trabalho, apenas os dois outros índices, o de Jevons e o de Dutot, são discutidos em mais detalhes[27].

2.3.2 – Os Índices de Dutot e de Jevons nos Contextos Axiomático e Estocástico

As fórmulas de Dutot e de Jevons, elaboradas, respectivamente, pelo francês Nicolas Dutot, em 1738,[28] e pelo economista inglês William Stanley Jevons, em 1863[29], são as mais adotadas por institutos oficiais de estatística para cálculos no nível elementar de análise.

A fórmula de Dutot, que representa uma razão de preços médios aritméticos entre dois períodos, é expressa matematicamente por:

[25] O teste da reversibilidade temporal assume que se os preços no período (t + 1) forem revertidos para os do tempo (t – 1), então o índice acumulado entre os periodos (t - 1) e (t + 1) deve ser igual a unidade. Ou seja, $I(p^{t-1}, p^t)I(p^t, p^{t+1}) = 1$

[26] Vale lembrar aqui o comentário de Fisher (1922, p. 29-30): ‘‘In fields other than index numbers it is often the best form of average to use. But we shall see that the simple arithmetic average produces one of the very worst of index numbers. And if this book has no other effect than to lead to the total abandonment of the simple arithmetic type of index number, it will have served a useful purpose.’’

[27] A propósito do uso dos índices e da importância de ponderadores adequados, Diewert (2003) comenta que os índices não-ponderados de Carli, Jevons e Dutot, amplamente usados no nível elementar, pecam por serem não-ponderados. Até recentemente, pensou-se que os vícios produzidos por índices não-ponderados fossem desprezíveis. Porém, com a disponibilidade cada vez maior de produtos com códigos de barra, foi possível calcular índices cada vez mais próximos dos índices ideais (por exemplo, Fisher, Walsh e Törnqvist Theil), o que levou à constatação de vícios fortemente significativos nos três índices tradicionalmente adotados. Ainda segundo Diewert (2003), as práticas empregadas pelos Institutos de Estatística para cálculo no nível elementar são, simplesmente, inconsistentes com os propósitos do índice de Laspeyres, o qual requer ponderadores adequados em todos os níveis de agregação.

[28] Velde (2009).

[29] Stanford Encyclopedia of Philosophy.

$$I_D\left(p^{t-1}, p^t\right) \equiv \frac{\sum_{k=1}^{K} \frac{1}{K} p_k^t}{\sum_{k=1}^{K} \frac{1}{K} p_k^{t-1}} = \frac{\sum_{k=1}^{K} p_k{}^t}{\sum_{k=1}^{K} p_k^{t-1}} \qquad (2.2)$$

onde:

$I_D\left(p^{t-1}, p^t\right)$ é o índice de Dutot entre os períodos t -1 e t

p_k^t é o preço do bem k no mês t

p_k^{t-1} é o preço do bem k no mês t-1

K é o número de bens específicos no agregado elementar

Alternativamente, a fórmula (2.2) pode ser transcrita como:

$$I_D\left(p^{t-1}, p^t\right) = \frac{\sum_{k=1}^{K} p_k^t}{\sum_{k=1}^{K} p_k^{t-1}} = \frac{\sum_{k=1}^{K} \left(\frac{p_k^t}{p_k^{t-1}}\right) p_k^{t-1}}{\sum_{k=1}^{K} p_k^{t-1}} = \sum_{k=1}^{K} \frac{p_k^t}{p_k^{t-1}} w_k^{t-1} \qquad (2.3)$$

onde $w_k = \frac{p_k^{t-1}}{\sum_{k=1}^{K} p_k^{t-1}}$ é o peso do produto *k* no tempo t-1. Esta equação indica que os produtos com preços mais altos têm um peso maior no resultado, o que não necessariamente reflete a importância do bem em termos de quantidade consumida pela população, o que implica vício no índice de Dutot quando aplicado a estratos não homogêneos.

Já a fórmula de Jevons, equivalente à média geométrica de razões de preços de K produtos entre dois períodos, é:

$$I_J\left(p^{t-1}, p^t\right) \equiv \prod_{k=1}^{K} \sqrt[K]{\frac{p_k^t}{p_k^{t-1}}} \qquad (2.4)$$

onde:

$I_J\left(p^{t-1}, p^t\right)$ é o índice de Jevons entre os períodos t -1 e t.

p_k^t é o preço do bem k no mês t.

p_k^{t-1} é o preço do bem k no mês t -1.

K é o número de bens específicos no agregado elementar

A principal argumentação econômica vinculada à justificativa da fórmula de Jevons é sua capacidade em captar o efeito substituição. Este índice, que tem por base curvas de indiferença do tipo Cobb-Douglas[30], assume que os consumidores mantêm despesas fixas e variam as quantidades. A fórmula de Jevons, do ponto de vista da abordagem axiomática, atende a todos os principais testes, dentre os quais o da reversibilidade temporal[31], o da circularidade[32], que estabelece que o índice acumulado em n períodos deve ser igual ao produto dos índices individuais e, o da comensurabilidade. Este último, que estabelece que o resultado do índice deve permanecer inalterado caso ocorram alterações nas unidades de medida dos produtos não é, entretanto, atendido pela fórmula de Dutot. Daí, a necessidade de agregados elementares homogêneos quando da utilização desta fórmula.

Na visão estocástica, os índices de Jevons (I_J) e de Dutot (I_D), sob a hipótese de variâncias constantes dos desvios dos preços em torno de suas médias em dois períodos são, teoricamente, bem próximos. Esta afirmação pode ser deduzida a partir da relação entre I_J e

[30] Diwert e Nakamura (1993, p. 48).
[31] Sobre o teste da reversibilidade temporal ver nota de rodapé [25].
[32] Para maiores detalhes consultar ILO et al. (2004, p.363)

I_D para os períodos (*t* – *1*) e (*t*), demonstrada por Diewert (1995, p. 27-28), que estabelece que:

$$I_J\left(p^{t-1}, p^t\right) = \prod_{k=1}^{K} \sqrt[K]{\frac{\overline{p}^t(1+e_k^t)}{\overline{p}^{t-1}(1+e_k^{t-1})}} = \frac{\overline{p}^t}{\overline{p}^{t-1}} \prod_{k=1}^{K} \sqrt[K]{\frac{(1+e_k^t)}{(1+e_k^{t-1})}} \qquad (2.5)$$

onde:

$\overline{p}^t$ é o preço médio dos K produtos em t.

$\overline{p}^{t-1}$ é o preço médio dos K produtos em t-1.

e_k^t e e_k^{t-1} são os desvios dos preços do produto k em relação às médias dos períodos t e t-1, respectivamente.

e, considerando que

$$\frac{\overline{p}^t}{\overline{p}^{t-1}} = I_D\left(p^{t-1}, p^t\right), \text{ vem:} \qquad (2.6)$$

$$I_J\left(p^{t-1}, p^t\right) = I_D\left(p^{t-1}, p^t\right) f\left(e^{t-1}, e^t\right), \qquad (2.7)$$

onde a função f é tal que

$$f(e^{t-1}, e^t) = \prod_{k=1}^{K} \sqrt[K]{\frac{(1+e_k^t)}{(1+e_k^{t-1})}} \qquad (2.8)$$

Se f for expandida em torno de $e^{t-1} = 0$ e de $e^{t} = 0$ por uma série de Taylor de segunda ordem, obtém-se a relação aproximada entre os dois índices como sendo:

$$I_J(p^{t-1}, p^t) \approx I_D(p^{t-1}, p^t)[1 + (1/2)\,\mathrm{var}(e^{t-1}) - (1/2)\,\mathrm{var}(e^t)] \qquad (2.9)$$

com $\mathrm{var}(e^{t-1})$ e $\mathrm{var}(e^t)$ representando, respectivamente, as variâncias dos desvios multiplicativos de preços em torno da média nos períodos (*t-1*) e (*t*).

Esta relação (2.9) é de grande relevância por permitir, como será visto no sétimo capítulo, a determinação do índice de Dutot com heterogeneidade controlada a partir do índice de Jevons com essa mesma característica. O índice de Jevons com heterogeneidade controlada pode ser obtido, por sua vez, a partir da modelagem dos preços, como é proposto neste trabalho.

Capítulo 3 – Não-resposta e Imputação: Aspectos Teóricos

Os painéis de preços nas pesquisas de índices de preços ao consumidor contêm vazios deixados pelos produtos não precificados. Se o painel representar um recorte de um período ou mesmo de alguns meses, o analista estará diante de um conjunto de dados denominado não balanceado. Para Singer e Willet (2004), nenhuma discussão sobre dados não balanceados deve ser realizada sem uma análise complementar sobre suas origens. Por isso, este terceiro capítulo destina-se ao comentário dos principais aspectos relacionados à ocorrência de dados faltantes em pesquisas de índices de preços ao consumidor, com ênfase em sua natureza e também nos cuidados que devem ser observados em métodos de tratamento da não-resposta e em procedimentos de imputação.

Como mencionado na introdução, os resultados dos IPCs precisam ser divulgados mensalmente após intenso processo de trabalho no qual enorme base de dados é construída, criticada e analisada. Na fase de coleta, são empregados procedimentos, em geral padronizados, para precificação dos produtos e serviços que compõem a cesta de consumo da população-objetivo dos indicadores. Entretanto, por mais eficiente que seja um modelo de coleta de preços e por mais regular que seja o comportamento do mercado em uma cidade, região ou país, é comum que muitos produtos que precisam ser pesquisados mensalmente não sejam encontrados nos balcões ou nas gôndolas dos estabelecimentos comerciais.

E podem ser inúmeras as causas dessas ausências: fornecedores podem ter problemas para entregar suas mercadorias; produtos podem sofrer alterações de embalagem ou gramatura, tornando-se em desconformidade com as especificações de coleta pré-determinadas; o produto pode ter sua produção interrompida por um fabricante por motivos de alto custo de insumos ou matérias primas; pode estar em falta por razões de irregularidades, como chuvas e geadas ou estar sob a influência de efeitos sazonais; o estabelecimento

comercial pode optar por não vender mais determinada mercadoria em decorrência de fraca demanda ou por estratégia comercial particular; o produto pode, até mesmo e “curiosamente”, estar ausente pelo fato de ser muito comercializado. É o caso, por exemplo, da falta de um bem motivada pela compra em excesso por parte da população, seja por preços baixos, seja por alta dos substitutos, seja por preço em oferta ou necessidades momentâneas aguçadas em virtude de exigências emergentes ou de moda. Outras causas de não-resposta podem estar atreladas a problemas operacionais, à recusa de informação, à impossibilidade de atendimento por parte do informante ou à falta de condições para visita a um estabelecimento comercial que se situe em área distante ou cujo acesso possa estar obstruído durante determinado período. Logo, a ausência da informação pode ser classificada em temporária ou permanente.

Para retratar as diversas possibilidades de ausências de preços que se apresentam quando da realização de suas pesquisas, uma vez que o tratamento a ser dado à não-resposta varia em decorrência de sua natureza, os institutos oficiais de estatística recorrem, em maior ou menor grau e de acordo com mecanismos internos de controle, a recursos diversos para tratar o problema.

Na prática, para identificação dos motivos das perdas, é comum que os responsáveis pela coleta de dados disponham de códigos que são atrelados aos preços coletados e auxiliam analistas nas justificativas das ausências de determinados produtos em alguns estabelecimentos comerciais.

Com relação às consequencias de se calcular estatísticas com base em conjuntos de dados incompletos, os analistas normalmente recorrem a critérios e métodos de imputação que têm por propósito minimizar os efeitos das perdas no resultado geral almejado. Por isso, aliada ao conhecimento da dinâmica do mercado, a compreensão dos fatores que resultam na ausência do preço ou do mecanismo gerador das perdas toma conotação crucial na escolha dos critérios e métodos a serem adotados.

Por exemplo, pode ser importante para o processo de identificação de critérios de imputação de preços distinguir a ausência relatada em campo em decorrência de fatores como interrupção de produção ou em função de aumento de demanda. No primeiro caso, a interrupção de produção poderia induzir o analista a investigar suas causas. Uma resposta do setor produtivo informando que um produto, uma marca de feijão, por exemplo, tem pouca procura, poderia remeter a um critério de imputação que prezasse pelo preço médio dos concorrentes, já que sendo o feijão um bem essencial para uma parcela considerável da população, seria aceitável supor que o consumidor, ao não encontrar a marca “A”, passasse a comprar a concorrente de preço semelhante, a marca “B”. Este procedimento, que substituiria a marca “A” pela marca “B” no cálculo do índice conduziria, pelo menos intuitivamente, a uma estatística representativa dos movimentos tanto dos preços praticados quanto dos pesos (importância) dos produtos efetivamente consumidos.

Por outro lado, se a marca de feijão “A” aparecesse na base de dados sem preço e se o motivo da ausência fosse identificado como sendo decorrente de redução momentânea de estoque em função de alta procura por parte dos consumidores, uma imputação pelo preço médio das cotações da mesma marca “A” encontradas nos demais estabelecimentos ou um critério que levasse em conta o preço desta marca “A” no mês anterior talvez pudessem ser considerados mais apropriados. Neste segundo exemplo, a não substituição da marca “A” seria adequado. Sob esta argumentação, depreende-se que atribuir a um bem esgotado em função de um número excessivo de unidades vendidas o preço de um bem substituto pouco comercializado pode gerar um vício no resultado do indicador.

Um outro exemplo ilustra como o método empregado pode também conduzir a estimativas viciadas: considere-se uma amostra aleatória de locais na qual seja predominante a existência de estabelecimentos de grande porte ou para a qual há grande ocorrência de dados faltantes para estabelecimentos de pequeno porte. Se os preços nos estabelecimentos de

grande porte forem recorrentemente menores que os praticados em locais de menor porte e, ainda, se parte da população realiza suas compras nos estabelecimentos de pequeno porte, os preços médios calculados por um modelo de médias aritméticas simples não serão estatísticas adequadas para representar o preço médio geral, pois, neste exemplo, a não-resposta, dependente de uma covariável, "porte", não pode ser considerada como completamente aleatória. Existe, aqui, um mecanismo gerador de perdas que não deve ser ignorado e a opção por métodos robustos de imputação, nestas circunstâncias, deve prevalecer (RUBIN, 1976). Ressalta-se, contudo, que a não-resposta não é, por sua simples ocorrência, um grave problema. De acordo com Singer e Willett (2003), as questões mais relevantes são as taxas com as quais ocorrem e o tipo de não-resposta observado. No caso dos índices de preços ao consumidor verificam-se altas taxas de não-resposta.

Como será discutido em termos formais nas seções adiante, desconsiderar o tipo de não-resposta e seu mecanismo gerador na hora de se escolher um método de imputação pode conduzir a estatísticas seriamente viciadas. É conveniente salientar inclusive que, dependendo do tipo de não-resposta presente, diversos métodos estatísticos de análise deixam de ter validade (RUBIN, 1976; LITTLE e RUBIN, 1987).

3.1 – Tipos de Não-resposta

Little e Rubin (1987) apresentaram uma sistematização do problema de não-resposta, elencando os tipos de dados faltantes que surgem nos levantamentos estatísticos. Explicações detalhadas sobre o tema podem também ser encontradas em Diggle, Liang e Zeger (1994), Schafer (1997) ou Little (1995). A partir desses estudos, é possível classificar a não-resposta como sendo informativa ou não informativa.

A não-resposta do tipo não informativa pode se enquadrar em completamente aleatória, cuja sigla recorrente na literatura é MCAR, do inglês *missing completely at random* ou aleatória, para a qual é comum a sigla MAR (missing at random).

A não-resposta completamente aleatória possibilita ao analista considerar que os casos observados de seu painel de dados constituem uma amostra aleatória simples do conjunto completo de dados, que incluiria todas as observações caso não houvesse não-resposta. Entretanto, a não-resposta do tipo MCAR é de rara ocorrência em levantamentos estatísticos. Para aceitar a suposição de que a não-resposta em um conjunto com dados faltantes é MCAR, a probabilidade de obtenção dos dados não pode estar associada nem aos preditores nem à variável resposta[33]. Qualquer potencial relação entre valores, observados ou não, da variável dependente e a probabilidade de não-reposta invalida esta suposição. No caso de dados longitudinais, a condição MCAR exige que não haja relação entre a probabilidade de obtenção de dados e o tempo. Transportando para os termos normalmente usados em pesquisas de IPC, esta última condição informa que a probabilidade de verificação do preço (p_i) de uma bem particular i não deve depender do mês (t) em que a pesquisa está sendo realizada. A não-resposta do tipo MCAR é altamente restritiva e requer o aceite de condições normalmente improváveis.

O outro tipo de não-resposta, MAR, é menos restritivo que MCAR, permitindo associação entre a probabilidade de ausência do preço e valores observados tanto da variável dependente quanto de preditores. Não se supõe, contudo, qualquer relação entre a probabilidade de não-resposta e valores não-observados da variável resposta ou de covariáveis.

[33] Esta segunda condição já impõe séria restrição à aceitação de MCAR, já que o conjunto completo da variável dependente é não observado.

3.2 – Abordagem Matemática

A não-resposta do caso MAR, levando-se em conta uma nomenclatura usual em pesquisas de índices de preços ao consumidor, pode ser enunciada seguindo-se a abordagem de Schafer (1997) da seguinte forma: considere-se um conjunto completo de dados representado por uma matriz Y (n x v) de preços composta por n produtos e v variáveis, onde p_i, por exemplo, pode fazer alusão a um vetor de preços mensais de um produto i, i = 1,...,n. Se for assumida a suposição de dados independentes e identicamente distribuídos (iid), a densidade de probabilidade do conjunto de dados pode ser aproximada por:

$$P(Y \mid \theta) = \prod_{i=1}^{n} f(p_i \mid \theta), \qquad (3.1)$$

onde f é a função de densidade de probabilidade para os preços de um produto particular e θ é um vetor de parâmetros. Assumindo-se que as distribuições f são multivariadas e dividindo-se o conjunto de dados Y em duas partes, Y_{obs} e Y_{falt}, onde Y_{obs} corresponde aos observados e Y_{falt} aos faltantes e, além disso, recorrendo-se a uma matriz (n x v) de variáveis indicadoras R, onde o elemento a_{iv} será 1 se o preço do produto correspondente i, dada a variável v, for observado e zero em caso contrário, pode-se organizar a correspondência matemática sobre os aspectos teóricos da não-resposta do tipo MAR da seguinte forma:

$$P(R \mid Y_{obs}, Y_{falt}, \xi) = P(R \mid Y_{obs}, \xi), \qquad (3.2)$$

onde ξ é um parâmetro do modelo gerador de perdas. Ou seja, se a distribuição de R condicionada ao conjunto completo de dados for idêntica à distribuição de R condicionada apenas aos preços efetivamente coletados, a não-resposta será aleatória (MAR).

Little e Rubin (1987) complementam o trabalho afirmando que, caso seja possível assumir que θ e ξ sejam distintos e que os dados são do tipo MAR, então o mecanismo de não-resposta será ignorável.

Schafer (1997, p. 12) demonstra que, diante de não-resposta ignorável, deixa de ser necessário conhecer o modelo gerador das perdas para realização de inferência com base em verossimilhança e que análises podem ser implementadas com a maximização do termo $P(Y_{obs} \mid \theta)$ da função de verossimilhança que se constrói quando se supõe não-resposta do tipo MAR. O autor referiu-se ao fator $P(Y_{obs} \mid \theta)$ como "*observed-data likelihood*" ou "verossimilhança para os dados obsevados".

Logo, diante de não-resposta do tipo MAR, deve-se avaliar com bastante cautela o uso de determinadas técnicas de inferência. Por exemplo, ignorar os dados ausentes e calcular médias apenas com os casos observados ou realizar imputações com base no último preço observado são técnicas inválidas sob a suposição deste tipo de não-resposta (RUBIN, 1976). Mesmo em amostras transversais, demonstra-se que diversos estimadores de médias são viciados. Os modelos longitudinais, que em geral recorrem a inferências baseadas em verossimilhanças, constituem, por sua vez, métodos adequados de análise (DEMIRTAS, 2004; LAIRD, 1988).

Nos capítulos cinco e sete, os modelos longitudinais hierárquicos são utilizados para avaliação do comportamento dos preços e para identificação de covariáveis capazes de explicar suas diferenças. No sexto capítulo, faz-se uma avaliação sobre os resultados de imputação gerados pelos modelos longitudinais e por três outros métodos frequentemente

adotados em institutos oficiais de estatística. A opção por se considerar os modelos longitudinais para uso em imputação deve-se, exatamente, ao fato de serem válidos mesmo sob a ocorrência de não-resposta do tipo MAR (Hedeker e Gibbons, 1997). Antes da modelagem, porém, faz-se, no próximo capítulo, um estudo descritivo dos preços para dados oriundo do SNIPC.

Capítulo 4 – Avaliação Descritiva no Contexto Longitudinal

Neste capítulo, após a apresentação dos dados disponíveis para a realização do trabalho, faz-se uma análise descritiva do crescimento dos preços dos produtos considerando-se seu caráter longitudinal. O estudo descritivo é um procedimento interessante para o trabalho por ser capaz de proporcionar subsídios para formulação de modelos e por contribuir para identificação de algumas características presentes nos dados, fornecendo "pistas" para respostas a questões gerais sobre mudanças nos preços de cada produto ao longo do tempo, ou nos diferentes padrões de evolução das séries de preços, bem como identificação de variáveis ou fatores "candidatos" a preditores para alterações nesses padrões. Para desenvolvimento do estudo e estimação dos modelos nos capítulos posteriores, serão utilizados dados do SNIPC. Como o escopo do trabalho limita-se ao nível mais desagregado, onde normalmente são aplicados os índices de Jevons e de Dutot, optou-se por se restringir o estudo a apenas um subitem[34]. Mais especificamente, as informações disponibilizadas correspondem às de cinco especificações[35] do subitem "arroz", comercializadas em estabelecimentos varejistas localizados na região metropolitana do Rio de Janeiro no período de outubro de 2006 a setembro de 2007[36]. Logo, o painel de preços constitui-se num "recorte" espacial, temporal e qualitativo do banco de dados do SNIPC. Nesta investigação descritiva, entretanto, serão considerados dados referentes a preços de apenas uma especificação.

[34] Supõe-se que os preços coletados sejam originários de um processo de amostragem aleatória simples.
[35] Especificação é a descrição detalhada do produto, de forma que possa ser identificado durante a coleta de preços.
[36] Por motivo de proteção a informações sigilosas, a condição estabelecida pelo IBGE para liberação do acesso às bases de dados do SNIPC para fins de análise nesta dissertação era a de que os produtos analisados não deveriam ser identificados.

4.1 – Base de Dados

O banco de dados é composto por 23 variáveis, englobando preços de cinco diferentes especificações de arroz e características associadas. São produtos comercializados em 74 estabelecimentos comerciais.

As variáveis utilizadas para modelagem, obtidas a partir do SNIPC são:

Bairro - *Bairro onde se localiza o estabelecimento comercial.*

Nível - *Variável categórica que indica a classificação socioeconômica do bairro em que o estabelecimento comercial está localizado.*

Tipo de local - *Classificação do tipo de estabelecimento comercial segundo o SNIPC. Por exemplo, os locais podem ser classificados como supermercados, mercearias, etc.*

Porte - *Variável categórica que classifica cada local como sendo de "grande porte" ou "pequeno/médio porte". Definida a partir do número de "pessoal ocupado" nas empresas, segundo o Cadastro Central de Empresas – CEMPRE/IBGE*[37].

Local - *Variável que determina o código do estabelecimento comercial onde o preço do produto é pesquisado.*

Produto – (especificação) *É a descrição do elemento amostral a ser pesquisado nos diferentes estabelecimentos comerciais. Por exemplo, o arroz da marca A, tipo I, longo fino, embalagem de 1kg constitui um produto.*

[37] Estatísticas (2008)

Produto-Local - *Esta variável determina de forma única a unidade amostral portadora da informação a ser coletada. Representa a associação do produto com o local em que é vendido. Por esta concepção, o arroz da marca A, tipo I, longo fino, embalagem de 1kg (como descrito anteriormente na variável "PRODUTO"), quando associado à variável "LOCAL", determina o elemento "produto-local". O mesmo produto em outro estabelecimento é um outro produto-local.*

Pessoal ocupado - *Faixa de pessoal ocupado, definida a partir do número de "pessoal ocupado" nas empresas, segundo o CEMPRE.*

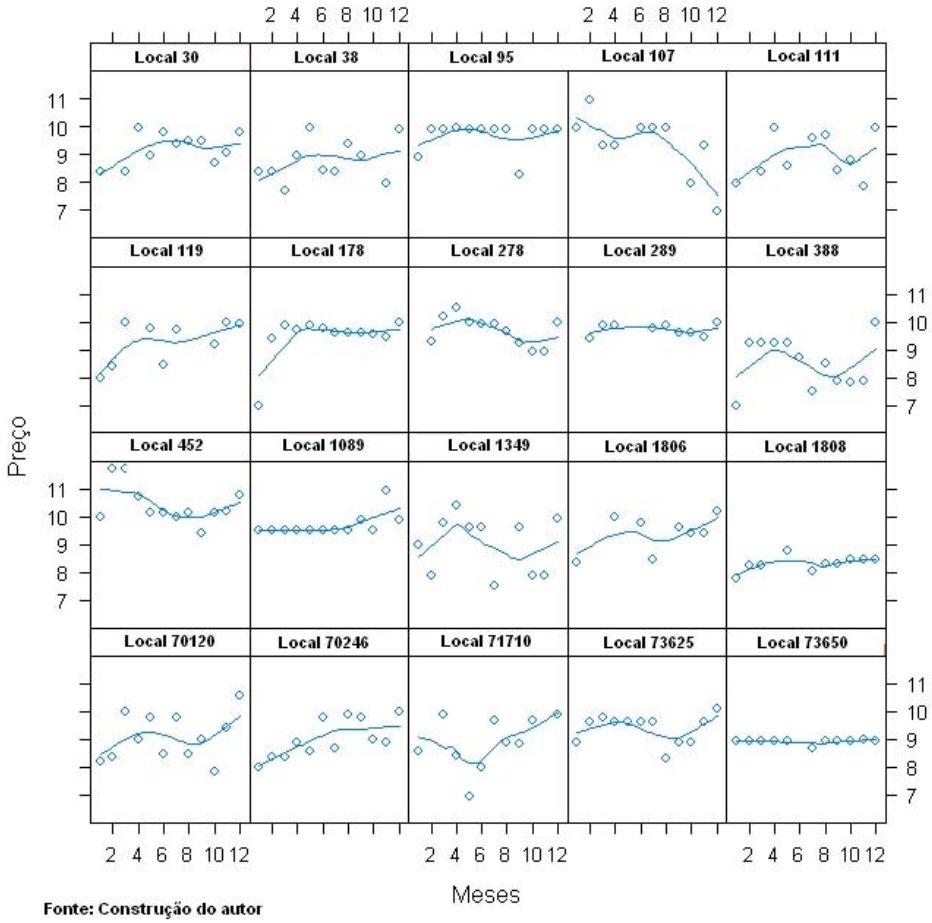
VAL1 a VAL12 - *Preços dos produtos pesquisados em cada local, em cada um dos 12 meses do período analisado.*

4.2 – Mudança Individual dos Preços ao Longo do Tempo

A observação do padrão de crescimento dos preços dos produtos comercializados nos estabelecimentos comerciais (produtos-locais) ao longo do tempo pode ser valiosa para a formulação da modelagem a ser realizada adiante. Uma análise simples para identificação e avaliação das trajetórias de crescimento dos preços pode ser obtida a partir da construção dos gráficos de curvas alisadas de forma não-paramétrica. Uma motivação para a abordagem não-paramétrica é a não exigência de suposições. A Figura 1 mostra gráficos com imposição de alisamento não-paramétrico para os preços praticados em 20 estabelecimentos comerciais

aleatoriamente selecionados[38] dos 74 cadastrados pelo SNIPC para o período de outubro de 2006 a setembro de 2007.

Figura 1: Alisamento não-paramétrico para os preços do produto 05000 praticados em 20 locais durante os meses de out/06 a set/07.



Fonte: Construção do autor

Na Figura 1, os estabelecimentos comerciais são identificados por códigos, por exemplo, “Local 30”. Com a observação dos gráficos com alisamento, percebe-se a evolução dos preços dos produtos ao longo de 12 meses. Alguns locais apresentam irregularidades nos

[38] Seleção aleatória simples sem reposição feita por intermédio da função “sample”, do R (R, 2008).

preços do produto *05000*, como os locais 388, 1349 e 71710. No local 107, o preço do arroz especificado cai de modo acentuado no período. Já nos locais 1808 e 73650, os preços apresentam baixíssima dispersão e uma modesta tendência de crescimento. Com exceção do local 73650, todos os demais parecem possuir curvas alisadas com um ou mais pontos de inflexão. Uma importante constatação é a de que o preço no último mês (set/07) está, em quase todos os 20 estabelecimentos, acima do preço do mês inicial (out/06). Nas curvas de alisamento essa característica também está presente, com as exceções ficando com os produtos comercializados nos locais 107, 278 e 452. Este local, o 452, ainda se destaca por sistematicamente ter preços acima da média global.

Focando-se mais pormenorizadamente o comportamento dos preços, parece que uma trajetória linear pode explicar o comportamento dos valores cobrados pelo produto *05000* nos locais 289, 1808 e 73650, enquanto que nos outros, trajetórias polinomiais talvez resultem em melhores ajustes.

A avaliação feita nesta seção, todavia, não leva em consideração nenhuma informação adicional sobre os produtos e locais como, por exemplo, região ou bairro em que se localiza o estabelecimento comercial, tamanho ou porte do mesmo, semana do mês em que o preço é coletado, etc. Estes dados podem ser importantes em modelos mais complexos. Nas seções seguintes, alguns desses preditores serão utilizados. A seguir, examinam-se os resultados de uma suavização paramétrica dos dados pelo método dos mínimos quadrados.

4.3 – Ajustes de Modelos por Local

Após a visualização de alisamento não-paramétrico feito na seção 4.2 para um conjunto selecionado de locais, o ajuste de modelos de regressão pelo método dos mínimos quadrados pode enriquecer a análise exploratória.

Apesar dos modelos de regressão por mínimos quadrados assumirem independência e homoscedasticidade dos resíduos, situações improváveis em estudos longitudinais, onde as perturbações de um produto em um determinado local ao longo do tempo tendem a ser autocorrelacionadas e heteroscedásticas, suas estimativas podem ser úteis se a finalidade for análise exploratória (Singer e Willett, 2003, p. 33).

O ajuste de trajetórias lineares aos preços dos 20 locais pode parecer, inicialmente, um tanto ineficiente, pois é natural que os dados de diferentes produtos se ajustem melhor a diferentes curvas. Entretanto, a regressão linear pode facilitar a investigação de diferenças nos locais selecionados quanto à prática de preços por intermédio de comparações dos parâmetros estimados para as regressões de cada estabelecimento.

A Figura 2 reforça o que foi verificado com o alisamento não-paramétrico dos preços com relação ao crescimento no período de outubro de 2006 a setembro de 2007 com a ilustração das retas de regressão estimadas para cada um dos 20 locais segundo a equação[39]

$$\begin{aligned} y_i &= \beta_0 + \beta_1 X_i + \varepsilon_i \\ \varepsilon_i &\sim N\left(0, \sigma^2\right) \end{aligned} \qquad (4.1)$$

onde:

y_i é o preço do produto 05000 no mês i ;

X_i assume valores de 1 a 12, de acordo com o mês i;

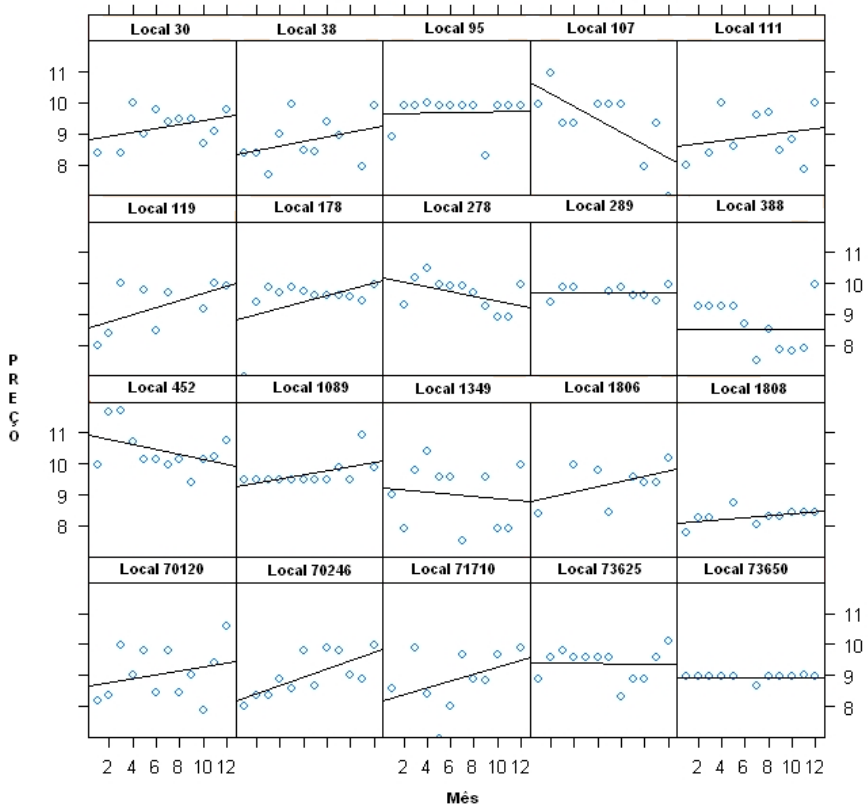
β_0 e β_1 representam o intercepto e coeficiente angular, respectivamente e;

ε_i é o termo de erro aleatório com distribuição $N\left(0, \sigma^2\right)$.

[39] Cada regressão envolve 12 observações.

Na maioria dos locais, a tendência é de alta. Alguns locais mantém preços com pouca oscilação em torno de um valor médio. No local 289, por exemplo, os preços flutuam próximos a R$9,70. Outros, apresentam elevada variabilidade, como os locais 111 e 70120.

Figura 2 – Ajuste de retas de regressão, via MQO, aos dados.



Fonte: Construção do autor com base em dados do SNIPC/IBGE.

Com o intuito de facilitar a análise, a Tabela 4.1 contém os valores estimados de intercepto, coeficiente angular, respectivos desvios-padrão, R^2 e variância residual para cada um dos 20 locais estudados, com o auxílio da função lm[40], do R[41].

É possível observar que os interceptos, representando preços praticados em outubro de 2006, têm valores estimados de R$ 8,08 a R$ 10,93. Quanto aos coeficientes angulares, suas

[40] Pinheiro e Bates (2000)

estimativas pontuais são distintas, apesar de muitas não serem estatisticamente diferentes de zero. Dos cinco locais com inclinação significativa, três apresentam inclinação positiva (locais 119, 1089 e 70246) e dois (locais 107 e 452), negativa.

Tabela 4.1 - Estatísticas das regressões aplicadas aos preços dos 20 locais selecionados

Local	*Preço estimado em out/06 (intercepto)*	*Desvio-padrão do intercepto*	*Coeficiente angular*	*Desvio-padrão do coeficiente angular*	*Variância residual*	R^2
30	8,78	0,38	0,06	0,05	0,30	0,15
38	8,32	0,45	0,07	0,06	0,53	0,12
95	9,63	0,34	0,00	0,04	0,30	0,00
107	10,68	0,55	-0,20	0,07	0,74	0,48
111	8,59	0,60	0,04	0,07	0,71	0,04
119	8,52	0,44	0,11	0,06	0,46	0,34
178	8,77	0,45	0,10	0,06	0,53	0,21
278	10,20	0,35	-0,07	0,04	0,23	0,23
289	9,70	0,17	0,00	0,02	0,05	0,00
388	8,50	0,58	0,00	0,08	0,90	0,00
452	10,93	0,40	-0,78	0,05	0,44	0,17
1089	9,23	0,22	0,06	0,03	0,13	0,32
1349	9,24	0,65	-0,03	0,09	1,14	0,02
1806	8,75	0,53	0,08	0,06	0,40	0,22
1808	8,08	0,15	0,03	0,02	0,05	0,23
70120	8,66	0,52	0,06	0,07	0,72	0,07
70246	8,14	0,31	0,13	0,04	0,25	0,49
71710	8,16	0,66	0,11	0,09	0,86	0,15
73625	9,38	0,32	0,00	0,04	0,28	0,00
73650	8,92	0,06	0,00	0,00	0,01	0,00

Fonte: Construção do autor

A última coluna da Tabela 4.1 mostra a estatística relacionada à qualidade do ajuste individual, R^2, que assume valores baixos, chegando a ter valor nulo para alguns locais, o que indica que modelos mais aprimorados serão necessários para representação satisfatória da evolução dos preços. A variância residual também oscila consideravelmente, retratando

[41] R Development Core Team (2008).

diferenças nos ajustes individuais. Para alguns locais, como os de códigos 38, 111, 388, 1349, 71710 e 70120, as retas de mínimos quadrados ajustam-se precariamente.

Esta avaliação preliminar pode subsidiar a modelagem longitudinal. Por exemplo, a indicação de que os interceptos e coeficientes angulares são distintos por local remete ao teste da hipótese de diferentes curvas de crescimento para explicação dos preços em diferentes estabelecimentos, o que poderá ser feito com a inclusão de coeficientes aleatórios nos modelos.

Numa investigação mais aprofundada, as trajetórias de crescimento poderão, inclusive, fugir ao padrão linear, como as estimadas pelo método dos mínimos quadrados, tornando-se polinomiais ou senoidais, dentre outras. Estas possibilidades serão consideradas no quinto capítulo, com a estimação dos modelos hierárquicos.

A propósito, a verificação de indícios sobre a adequabilidade de consideração de diferentes curvas ao invés de um modelo tradicional para explicação dos preços, curvas essas que seriam expressas por meio de um modelo hierárquico longitudinal com coeficientes aleatórios, pode ser iniciada com a observação de algumas estatísticas das estimativas dos parâmetros dos modelos de regressão linear simples até então estimados, como interceptos e coeficientes angulares. A análise simples de médias e desvios-padrão dessas estimativas pode auxiliar as resposta às perguntas básicas que pautam o estudo longitudinal que aqui se realiza e que são: a) *os níveis de preços são significativamente diferentes*? b) *como os preços dos produtos mudam com o tempo?*

É exatamente o que se pode inferir com a análise de interceptos e coeficientes angulares. A Tabela 4.2, a seguir, mostra a média, o desvio-padrão e a variância das estimativas para os dois parâmetros:

Tabela 4.2 - Estimativas dos parâmetros das regressões por mínimos quadrados.

Parâmetro	*Média*	*Desvio-padrão*	*Variância*	*CV*
Intercepto	9,06	0,81	0,67	8,90%
Coef. angular	0,025	0,081	0,067	324%

Fonte: Construção do autor

Estes dois parâmetros que integram a Tabela 4.2 são, obviamente, suficientes para determinação de uma reta. Entretanto, uma reta construída com seus valores médios seria representativa do "local médio" (simbólico) do conjunto de locais pesquisados. Porém, a informação mais importante em termos longitudinais é que interceptos e coeficientes angulares individuais, com valores diferentes de local para local, representam curvas particulares de evolução dos preços. Com base na amostra estudada de 20 locais, o intercepto médio teve valor igual a R$9,06, representando o preço médio praticado nos locais no mês de outubro de 2006, com desvio-padrão de R$0,81 e crescimento mensal médio de R$0,025. Este crescimento mensal, expresso pelo coeficiente angular, difere consideravelmente de um local para outro, o que leva ao altíssimo coeficiente de variação (324%, na quinta coluna da tabela 4.2) e reforça o já comentado sobre a necessidade de avaliação de um modelo que, formalmente, preze por essas diferenças.

Outra estatística de interesse é a correlação entre intercepto e coeficiente angular. No caso da amostra pesquisada, esta correlação assumiu o valor de -0,82, indicando que produtos que em outubro de 2006 tinham preços mais elevados, apresentaram menores taxas de mudança ao longo dos meses. Em outras palavras, locais com preços iniciais mais elevados mantiveram alterações menos expressivas nos níveis de preços de um mês para outro.

As diferenças nas curvas de crescimento dos preços, por sua vez, podem, ainda, decorrer da vinculação entre locais ou produtos com determinados preditores. Por exemplo, é possível que locais com preços mais baixos sejam de estabelecimentos de grandes redes varejistas, ou de maior porte. Ou ainda, que estejam localizados em bairros de menor poder aquisitivo por parte da população. Essa avaliação é feita a seguir.

4.4 – Avaliação Preliminar de Potenciais Fontes de Heterogeneidade dos Preços

Hawkes e Piotrowski (2003) observaram que um agregado elementar deve compreender, pelo menos, quatro dimensões: *tempo, produto, espaço e setor*. Estas dimensões, que podem ser percebidas como fontes de variância ou heterogeneidade dos preços, devem ser levadas em consideração em diversas etapas do IPC, por exemplo, para determinação dos agregados elementares, para dimensionamento amostral, para análise de resultados e até para percepção de como a dispersão dos preços pode influenciar resultados dependendo da fórmula de cálculo adotada, como analisado no sétimo capítulo em relação à fórmula de Dutot.

Neste trabalho, o domínio de análise sobre o qual o estudo se desenvolve hipotetiza um agregado elementar de acordo com estas dimensões: existe uma delimitação temporal, dada pelo período de coleta de preços, que é mensal; existe uma identificação de produto, já que a amostra é formada por marcas de um mesmo subitem e há também uma delimitação espacial, a cidade do Rio de Janeiro. A quarta dimensão, a setorial, está vinculada à atividade econômica. A princípio, percebe-se que, uma vez que os preços aqui estudados são provenientes de estabelecimentos comerciais varejistas, esta dimensão também é contemplada. Entretanto, ainda que os agregados elementares possam ser construídos com este embasamento teórico, as suposições a respeito do grau de homogeneidade devem ser

investigadas mais pormenorizadamente, já que, mesmo em estratos aparentemente homogêneos, há a possibilidade de influências de fatores indutores de heterogeneidade não considerados num primeiro momento, como subáreas dentro de uma região ou uma subclassificação de empresas em relação a volume de vendas que, uma vez identificados, podem ser úteis para o entendimento sobre alterações nos níveis de preços.

No caso brasileiro, a Classificação Nacional da Atividade Econômica, CNAE, que classifica as empresas de acordo com a atividade econômica a qual pertencem e que, em níveis mais desagregados possibilita, por exemplo, diferenciar grandes redes varejistas de estabelecimentos de menor porte, pode ser uma excelente fonte de informação para estratificação da amostra em IPC.

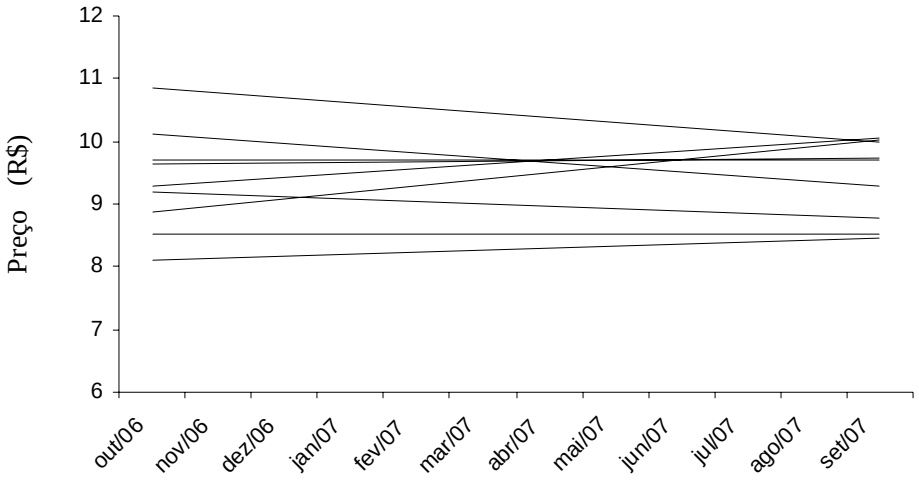
Logo, introduz-se na avaliação dois preditores relacionados aos estabelecimentos comerciais: *Porte (P)*, variável binária construída a partir de informações sobre o número de pessoal ocupado das empresas, que as classificam em pequenas ou médias (P=0) ou grandes (P=1); e *Nivel (N)*, variável dicotômica que indica a classificação socioeconômica do bairro onde se localiza, assumindo valores baixo ou médio (N=0) e alto (N=1)[42].

As Figuras 3 e 4 representam os ajustes das retas de regressão para os locais de pequeno/médio porte e grande porte, respectivamente.

Nos locais de porte pequeno ou médio o comportamento parece mais heterogêneo, com alguns preços sofrendo incrementos no período e outros sofrendo reduções, sempre dentro de patamares compreendidos entre R$8,00 e R$11,00.

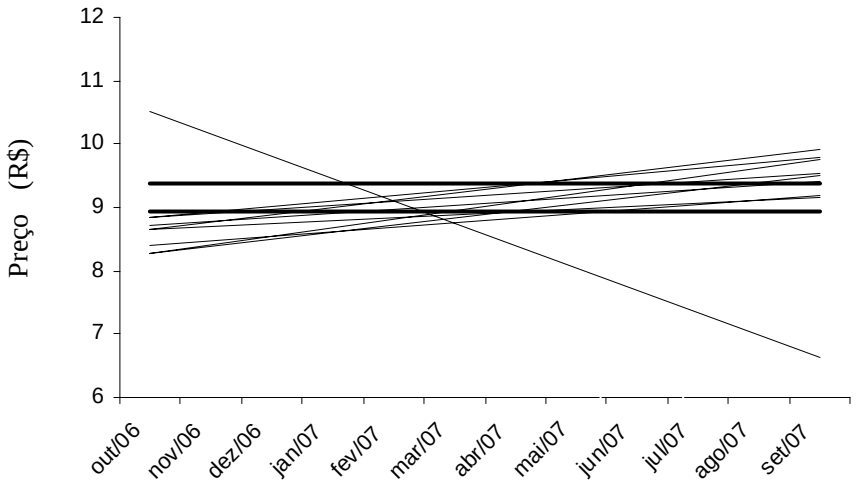
[42] Obtidas conforme descrição apresentada na seção 4.1.

Figura 3 – Retas ajustadas, por mínimos quadrados, para estabelecimentos de pequeno/médio porte.



Fonte: Construção do autor

Figura 4 – Retas ajustadas, por mínimos quadrados, para estabelecimentos de grande porte.

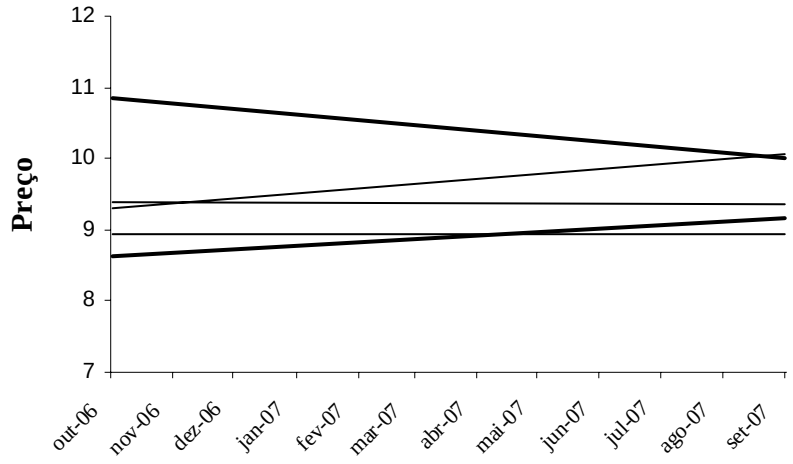


Fonte: Construção do autor

Os gráficos ilustram indícios de diferenças interessantes no comportamento dos preços. Os locais de grande porte, por outro lado, são mais homogêneos quanto ao comportamento de aumento de preços. Com exceção de um dos locais, identificado como sendo o local 107, as regressões apontam para reajustes regulares e similares entre outubro de 2006 e setembro de 2007.

Quando os estabelecimentos são agrupados segundo a condição socioeconômica do bairro em que estão localizados[43], observa-se que os preços dos produtos selecionados praticados nos bairros de maior poder aquisitivo sofrem incrementos, com exceção para os preços de um dos locais, cuja trajetória é decrescente (Figura 5).

Figura 5 - Retas ajustadas, por mínimos quadrados, para estabelecimentos localizados em bairros de nível socioeconômico alto.

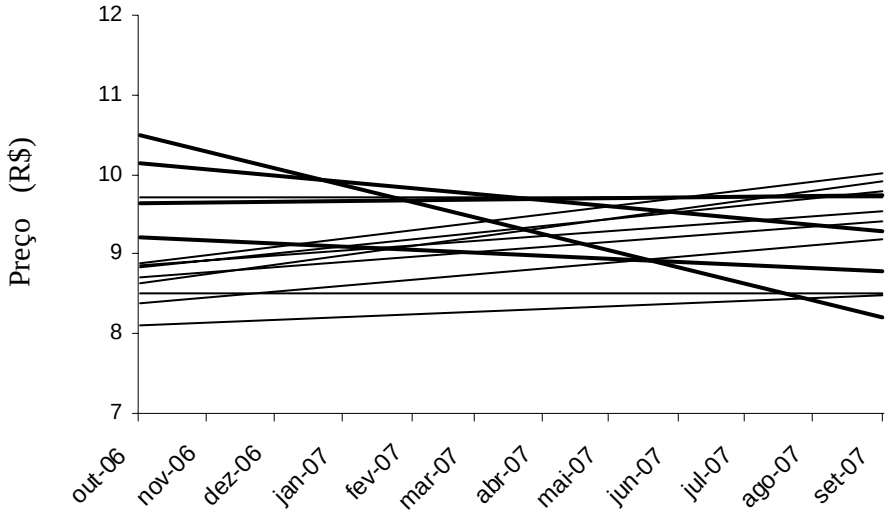


Fonte: Constução do autor

[43] Esta classificação socioeconômica, extraída do banco de dados do SNIPC, foi estabelecida por analistas do IBGE.

Este local que foge à regra, com redução de preços no período, praticava valores em outubro de 2006 consideravelmente acima dos demais, o que pode ter forçado sua trajetória a convergir para a média em setembro de 2007[44].

Figura 6 - Retas ajustadas, por mínimos quadrados, para estabelecimentos localizados em bairros de nível socioeconômico baixo ou médio.



Fonte: Construção do autor

Na Figura 6, que contém as retas de mínimos quadrados para os locais de bairros classificados como sendo de baixo/médio poder aquisitivo segundo o SNIPC, são observados aumentos de preços para a maioria dos locais analisados. Como no caso dos locais em bairros de alto poder aquisitivo, alguns dos estabelecimentos localizados nestas áreas também apresentaram redução de preços.

A análise dos dois preditores realizada nesta seção revela padrões mais claros de diferenças nos preços em função de diferenças no porte dos estabelecimentos. Quando o nível

[44] Sousa e Canedo (1999) identificaram os preços da concorrência como um dos fatores determinantes para formação dos preços em supermercados.

socioeconômico do bairro é avaliado, entretanto, as diferenças não aparecem de forma límpida.

A estimação de modelos mais complexos em etapa futura permitirá confirmar a importância desses fatores no nível e variação dos preços. Nos próximos capítulos, é realizada a inferência com base em modelos longitudinais capazes de contabilizar, durante o processo de estimação, a estrutura de covariância presente nos dados.

Capítulo 5 – Modelos Hierárquicos Longitudinais

Os modelos lineares, como os adotados no capítulo anterior para regressão da variável "preço" na variável "tempo", são estimados habitualmente pelo método dos mínimos quadrados ordinários (MQO). Estes modelos apresentam, dentre suas características, a suposição de erros normais, independentes, identicamente distribuídos (*iid*) e com variância constante $\left(\sigma^2\right)$. Matematicamente, esta abordagem pode ser representada pela seguinte formulação:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij} \qquad (5.1)$$

$$\varepsilon_{ij} \sim N\left(0, \sigma^2\right)$$

onde y_{ij} representa a resposta do indivíduo i (i=1,...,N) no tempo t_{ij} (j = 1,2, ... , n), sendo que esta variável "tempo" pode fazer alusão a qualquer unidade de medida temporal, como meses, semanas, dias, etc.

Para a maior parte dos estudos longitudinais, porém, estas suposições apresentadas para a distribuição dos erros são um tanto irrealistas. Além disso, em dados de painel[45] é plausível a expectativa de que as observações tomadas numa mesma unidade amostral tenham algum grau de dependência. O modelo (5.1) não considera a estrutura de variância e correlação eventualmente presente nos erros nem a estrutura de covariância da variável dependente y. Outro atributo deste modelo é que a tendência média estimada ao longo do tempo é a mesma para qualquer unidade amostral, já que os parâmetros β_0 e β_1 são fixos.

Os modelos de efeitos mistos (MEMs), usados neste trabalho para a modelagem de preços de produtos comercializados em estabelecimentos comerciais, apresentam algumas

vantagens se comparados aos modelos de regressão padrão. A principal delas é a possibilidade de inclusão de efeitos aleatórios individuais em suas equações. As estimativas de coeficientes individuais, como interceptos e coeficientes angulares, permitem o traçado de trajetórias específicas para cada unidade amostral, ampliando o horizonte de análise. No caso dos dados de IPC, por exemplo, estas trajetórias podem ser percebidas como estimativas das curvas de evolução dos preços dos produtos ao longo dos meses. Os efeitos aleatórios presentes nos modelos podem captar, ainda, a influência de cada indivíduo em seus respectivos valores observados. Em outras palavras, os MEMs constituem-se numa interessante ferramenta para avaliação da dependência entre as medidas de uma mesma unidade amostral.

Remetendo ao contexto dos índices de preços ao consumidor, e lembrando que o escopo do problema, num primeiro momento, está limitado a uma amostra de produtos de mesma especificação[46], um modelo com coeficientes aleatórios que capte a influência de cada estabelecimento (i) nos preços coletados ao longo dos meses pode ser obtido a partir da extensão da equação (5.1) com a inclusão de uma componente aleatória. Este modelo pode ser expresso como:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \zeta_{0i} + \varepsilon_{ij} \qquad (5.2)$$

$$\zeta_{0i} \sim N(0, \sigma_0^2)$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

[45] Denominação comum na literatura para conjuntos de medidas repetidas no tempo das mesmas unidades de análise (Singer, 2005).

[46] No sexto capítulo a modelagem é ampliada para uma amostra com diferentes marcas de arroz. Faz-se aqui, neste primeiro momento, a análise de um produto específico (marca) porque os métodos de imputação, com os quais se pretende comparar os resultados da modelagem, são aplicáveis, tanto no Brasil quanto em alguns outros países, a estratos homogêneos.

Em (5.2), ζ_{0i} representa a influência do estabelecimento *i* nos preços coletados. Como se supõe que os locais são uma amostra aleatória da população de estabelecimentos comerciais para o estrato analisado, os efeitos ζ_{0i} são considerados como provenientes de uma distribuição normal. Condicionados a esses efeitos aleatórios individuais ζ_{0i}, os erros ε_{ij} são independentes e normalmente distribuídos (HEDEKER e GIBBONS, 2006). Observa-se, entretanto, que o efeito do tempo no modelo (5.2) permanece fixo.

Flexibilizando mais ainda a equação (5.2), a introdução de um coeficiente angular aleatório pode permitir a avaliação mais aprimorada de trajetórias individuais de evolução dos preços nos estabelecimentos. Neste caso, a representação passa a ser:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \zeta_{0i} + \zeta_{1i} t_{ij} + \varepsilon_{ij} \qquad (5.3)$$

$$\begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{\mathbf{01}} \\ \sigma_{\mathbf{10}} & \sigma_{\mathbf{1}}^2 \end{bmatrix} \right)$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

onde:

y_{ij} é o preço do produto do local *i* no mês *j*;

β_0 é o intercepto global;

ζ_{oi} é o efeito aleatório que representa o desvio do intercepto do locai i em relação ao intercepto global;

β_1 é o coeficiente angular global;

ζ_{1i} é o efeito aleatório que representa o desvio do coeficiente angular do locai *i* em relação ao coeficiente angular global;

σ_0^2 representa a variância dos interceptos individuais, ζ_{0i};

σ_1^2 representa a variância dos coeficientes angulares individuais, ζ_{1i};

σ_{01} é a covariância entre os efeitos aleatórios e;

ε_{ij} é o erro aleatório.

Por permitirem a análise flutuar em dois terrenos, o intraindividual e o interindividual, os MEMs podem ser apresentados de acordo com uma configuração que explicite estes níveis hierárquicos. Assim, o primeiro nível de análise (nível I), conhecido como modelo de crescimento individual (SINGER e WILLETT, 2003), deve representar a mudança esperada nos preços de cada estabelecimento comercial. O segundo nível conterá equações que representam as curvas de crescimento que, estabelecidas através de seus respectivos parâmetros, interceptos e coeficientes angulares, permitem avaliar suas diferenças em função de preditores selecionados. O modelo (5.3), reescrito de acordo com este formato hierárquico mais evidente, torna-se:

$$\begin{aligned} y_{ij} &= b_{0i} + b_{1i}t_{ij} + \varepsilon_{ij} \qquad \text{(nível I)} \\ b_{0i} &= \beta_0 + \zeta_{0i} \qquad \text{(nível II)} \\ b_{1i} &= \beta_1 + \zeta_{1i}\,, \end{aligned} \tag{5.3.1}$$

$$\begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{\mathbf{01}} \\ \sigma_{\mathbf{10}} & \sigma_{\mathbf{1}}^2 \end{bmatrix} \right)$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

onde:

b_{oi} é o intercepto do local *i*;

b_{1i} é o coeficiente angular do estabelecimento comercial *i*;

e os demais parâmetros são todos como em 5.3.

O formato de apresentação dos modelos (5.2) e (5.3) é chamado por Singer e Willett (2003) de forma composta, enquanto que o formato (5.3.1) é denominado de multinível ou hierárquico. Vale observar que estes formatos são equivalentes: por exemplo, a substituição das subequações do modelo (5.3.1) na equação principal (equação do nível I, intraindividual) leva ao formato composto (5.3).

Esta construção hierárquica mostra como os MEMs permitem a inspeção dos processos de mudança nos preços em dois âmbitos de análise: no primeiro, focando as características das séries de cada estabelecimento comercial; no segundo, possibilitando o conhecimento das diferenças interindividuais ou, em outras palavras, permitindo identificar como os preços experimentam respostas distintas sob a influência de determinados preditores. A inclusão dos preditores pode ser feita facilmente em ambas as formulações.

5.1 – Notação Matricial

Uma terceira opção de apresentação, adotada em algumas etapas na seqüência do trabalho, é a matricial. Seja y_i um vetor $n_i x 1$ de preços do estabelecimento comercial i. Um modelo de regressão de efeitos mistos pode ser escrito, neste caso, como:

$$\underset{n_i x 1}{y_i} = \underset{n_i x p}{X_i}\ \underset{p \times 1}{\beta} + \underset{n_i x\ r}{Z_i}\ \underset{r \times 1}{\zeta_i} + \underset{n_i \times 1}{\varepsilon_i} \qquad (5.4)$$

$$\varepsilon_i \sim N\left(0, \sigma^2 I_{n_i}\right)$$
$$\zeta_i \sim N\left(0, \Sigma_\zeta\right)$$

com $i = 1...N$ estabelecimentos, $j = 1...n_i$ preços relacionados ao local i. X_i é a matriz de covariáveis $(n_i \; x \; p)$ do estabelecimento *i*, β é o vetor $(p \, x \, 1)$ de parâmetros fixos, Z_i é a matriz $(n_i \; x \, r)$ de desenho dos efeitos aleatórios, ζ_i é o vetor $(r \, \mathrm{x} \, 1)$ de efeitos aleatórios e ε_i é o vetor de erros $(n_i \, \mathrm{x} \, 1)$; I_{n_i} é a matriz identidade e Σ_ζ é a matriz de covariância $(r \, x \, r)$ dos efeitos aleatórios.

A distribuição conjunta das observações y_i e dos efeitos aleatórios ζ_i é normal multivariada, dada por:

$$\begin{bmatrix} y_i \\ \zeta_i \end{bmatrix} \sim N\left(\begin{bmatrix} X_i\beta \\ 0 \end{bmatrix}, \begin{bmatrix} Z_i\Sigma_\zeta Z' + \sigma^2 I_{n_i} & Z_i\Sigma_\zeta \\ \Sigma_\zeta Z'_i & \Sigma_\zeta \end{bmatrix} \right) \qquad (5.5)$$

Portanto, a matriz de variância e covariância das observações y é

$$V(y_i) = \Sigma = Z_i\Sigma_\zeta Z'_i + \sigma^2 I_{n_i} \qquad (5.6)$$

composição esta que demonstra como a suposição a respeito dos efeitos aleatórios influi nas estimativas das componentes das estruturas de variância e covariância de y.

5.2 – Modelos Incondicionais de Média e Crescimento de Preços

Um importante parâmetro a ser estimado na fase inicial de modelagem dos preços é o coeficiente de correlação intraclasse (ρ). O interesse nesta estatística decorre do fato de que seu valor representa a parcela da variabilidade total que é atribuída a diferenças entre os estabelecimentos comerciais.

A variância do nível I, intragrupo ou intralocal, por exemplo, só pode ser reduzida com a inclusão de preditores variantes no tempo. Já os preditores invariantes, como a

classificação do estabelecimento comercial como sendo de pequeno ou grande porte, reduz apenas a parcela de variância observada *entre* os estabelecimentos (variância interindividual). Os modelos podem ser adequados ou não em decorrência de proporcionarem a redução da variabilidade nessas duas dimensões.

O coeficiente de correlação intraclasse, cuja fórmula é:

$$\rho = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_\varepsilon^2} \qquad (5.7)$$

pode ser obtido a partir de um modelo que contenha apenas o intercepto aleatório, sem qualquer outro preditor. Este modelo, conhecido como *modelo de média incondicional*, proporciona a separação da variabilidade total dos preços nas duas componentes de variância e o cálculo de $\hat{\rho}$. Sua expressão na versão hierárquica é:

$$\begin{aligned} y_{ij} &= b_{0i} + \varepsilon_{ij} \\ b_{0i} &= \beta_0 + \zeta_{0i} \\ \varepsilon_{ij} \sim N\left(0, \sigma_\varepsilon^2\right) &\text{ e } \zeta_{0i} \sim N\left(0, \sigma_0^2\right) \end{aligned} \qquad (5.8)$$

onde:

y_{ij} é o preço do produto-local *i* no mês *j*;

b_{oi} é a média de preço do produto-local *i*;

β_0 é o preço médio global;

ζ_{oi} representa o desvio do intercepto do locai *i* em relação ao intercepto global;

σ_0^2 é a variância do preço médio dos produtos-locais;

σ_ε^2 é a variância de ε_{ij}, o erro aleatório do nível I;

O ajuste do modelo (5.8), por máxima verossimilhança (MV), aos preços do produto de código *05000* pesquisados nos 74 estabelecimentos comerciais que compõem a amostra, com a utilização da função *lme*[47], do R, está sintetizado na Tabela 5.1.

Tabela 5.1 – Resultado do modelo de média incondicional (5.8).

Parâmetro / estatística	*Estimativa*
β_0	9,26
σ_ε^2	0,51
σ_0^2	0,22
AIC	1834

Fonte: Construção do autor

Os valores apontam para um preço médio global estimado de R$9,26. A variância intralocal $\left(\sigma_\varepsilon^2\right)$ é de 0,51 e a entre locais $\left(\sigma_0^2\right)$ de 0,22. Nota-se que a variabilidade dentro dos estabelecimentos comerciais é maior que a variância entre locais.

A estimativa de ρ, $\hat{\rho}$, com a aplicação de (5.7), é 0,30, valor que indica que cerca de 30% da variância dos dados decorre de diferenças entre estabelecimentos comerciais e que 70% da variabilidade corresponde a fatores intraindividuais ou, em outras palavras, inerentes aos estabelecimentos. Estes fatores podem ser, por exemplo, estoques, que variam de um mês para outro ou o volume de vendas, dentre outros.

Para a redução da variabilidade não explicada em cada local, a inclusão da variável "*t*", representando o mês da coleta, é outro importante elemento. Sua inclusão, retratando a

[47] Pinheiro et al. (2008)

evolução dos preços de outubro de 2006 a setembro de 2007, é o caminho natural a ser seguido.

Um novo modelo, incluindo em princípio apenas este preditor temporal, pode ser escrito como:

$$\begin{aligned} y_{ij} &= b_{0i} + b_{1i}(t_{ij}) + \varepsilon_{ij} \\ b_{0i} &= \beta_0 + \zeta_{0i} \\ b_{1i} &= \beta_1 + \zeta_{1i}\,, \end{aligned} \qquad (5.9)$$

$$\varepsilon_{ij} \sim N\left(0, \sigma_{\varepsilon}^2\right)$$

$$\begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{\mathbf{01}} \\ \sigma_{\mathbf{10}} & \sigma_{\mathbf{1}}^2 \end{bmatrix}\right)$$

O modelo (5.9) é conhecido como modelo de crescimento incondicional. No formato composto assume a seguinte configuração:

$$y_{ij} = \beta_0 + \zeta_{0i} + \beta_1(t_{ij}) + \zeta_{1i}(t_{ij}) + \varepsilon_{ij} \qquad (5.10)$$

$$\varepsilon_{ij} \sim N\left(0, \sigma_{\varepsilon}^2\right)$$

$$\begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix}\right)$$

onde:

y_{ij} é o preço do produto-local *i* no mês *j*;

b_{oi} é o preço médio do produto-local *i*;

b_{1i} é o coeficiente angular da trajetória linear de crescimento do preço no estabelecimento comercial *i*;

β_0 é o preço médio geral em outubro de 2006;

β_1 é a taxa média de crescimento dos preços;

σ_0^2 representa a variância dos interceptos individuais, ζ_{0i};

σ_1^2 representa a variância dos coeficientes angulares individuais, ζ_{1i} e;

ε_{ij} é o erro aleatório do nível I.

A estimação do modelo (5.10), que assume trajetórias lineares para o crescimento dos preços dos produtos nos locais leva, de fato, à redução da variabilidade não explicada dentro dos estabelecimentos, como depreende-se da Tabela 5.2.

Tabela 5.2 – Resultado do modelo de crescimento incondicional 5.10, estimado por MV.

Parâmetro / estatística	*Estimativa*
σ_ε^2	0,47
σ_0^2	0,41
σ_1^2	0,002
ρ_{01}	-0,71
AIC	1814

Fonte: Construção do autor

De acordo com a suposição teórica, a inclusão da variável "*t*" provocou, como esperado, redução da parcela da variância intragrupo, σ_ε^2, cujo valor decresceu em cerca de 7,8%, de 0,51 para 0,47. Vale lembrar que σ_ε^2, neste modelo, faz alusão à dispersão dos preços ao redor das retas de crescimento estimadas para cada estabelecimento comercial.

A variância dos interceptos individuais, σ_0^2, foi de 0,41, enquanto que σ_1^2, a variabilidade observada nos coeficientes angulares das trajetórias de preços foi de 0,002. Entretanto, como houve inclusão de uma variável explicativa ("*t*") na equação intralocal, essas variâncias não devem ser diretamente comparadas com as do modelo anterior. Servem, contudo, de base para comparações com os valores assumidos pelas componentes de variância com a inclusão de preditores invariantes no tempo.

A análise do modelo de crescimento incondicional não deve prescindir, porém, da avaliação da correlação entre intercepto e coeficiente angular, ρ_{01}. Além de retratar a relação entre os resíduos do segundo nível, ζ_{0i} e ζ_{1i}, este parâmetro fornece "pistas" sobre a relação entre os preços no início do período (outubro de 2006) e a taxa de crescimento dos mesmos, oferecendo indícios sobre o comportamento geral nos 12 meses pesquisados.

A correlação estimada entre os dois parâmetros foi -0,71. Por este resultado, é possível depreender que locais que apresentaram preços mais altos em outubro de 2006 tiveram uma evolução do nível de preços mais moderada que os locais com preço mais baixos neste mês.

O critério AIC[48] (1814) do modelo (5.10), que tem o tempo como variável explicativa e que conta também com coeficientes aleatórios para intercepto e coeficiente angular foi, como esperado, menor que o do modelo de média incondicional (1834). O teste da razão de verossimilhança entre os modelos (5.8) e (5.10) foi 26,11, com *p-valor* inferior a 0,0001, ressaltando a melhor performance do segundo.

Nas seções seguintes, potenciais preditores teoricamente importantes para explicação do comportamento dos preços, identificados na avaliação feita no quarto capítulo, são incorporados na análise.

[48] Akaike Information Criterium (AKAIKE, 1973).

5.3 – Avaliação da Variabilidade entre Estabelecimentos

A inclusão de preditores nas equações que têm como variáveis respostas o intercepto e o coeficiente angular, permite retratar a influência das variáveis explicativas na redução da variância das estimativas dos parâmetros. A importância de cada um dos preditores pode ser verificada com a estimação de modelos como o (5.11), a seguir, uma extensão do modelo de crescimento incondicional.

No caso específico de avaliação da influência do porte do local na redução de σ_0^2 e σ_1^2, as equações a serem estimadas devem contemplar esta variável *Porte (P)* e seus coeficientes, apresentando o seguinte aspecto:

$$\begin{aligned} y_{ij} &= b_{0i} + b_{1i}(t_{ij}) + +\varepsilon_{ij} \\ b_{0i} &= \beta_0 + \beta_{02} P_i + \zeta_{0i} \\ b_{1i} &= \beta_1 + \beta_{12} P_i + \zeta_{1i} \end{aligned} \qquad (5.11)$$

$$\varepsilon_{ij} \sim N\left(0, \sigma_\varepsilon^2 I\right)$$

$$\begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{\mathbf{01}} \\ \sigma_{\mathbf{10}} & \sigma_{\mathbf{1}}^2 \end{bmatrix} \right)$$

onde:

β_{02} e β_{12} representam, respectivamente, as influências do preditor "Porte" no intercepto b_{0i} e coeficiente angular b_{1i};

σ_0^2 representa a variância de ζ_{0i};

σ_1^2 a variância de ζ_{1i};

σ_{01} é a covariância entre ζ_{0i} e ζ_{1i};

ε_{ij} é o erro aleatório do nível I, com variância σ_{ε}^2.

A investigação sobre a importância do nível socioeconômico do bairro (variável *Nível*) pode ser realizada de forma semelhante.

A Tabela 5.3, com os *pseudo-R^2* computados para as componentes de variância σ_0^2 e σ_1^2, contém os efeitos nas variabilidades do intercepto e coeficiente angular provocados pela inclusão de cada uma das duas variáveis no modelo de crescimento incondicional, como em (5.11).

Tabela 5.3 - Estatísticas pseudo-R^2 para a redução da variância entre locais com a inclusão das variáveis "Porte" e "Nível".

Pseudo-R^2	*Variável*	
	Porte	*Nível*
$R^2_{\zeta_0}$	0,20	0,00
$R^2_{\zeta_1}$	0,10	0,00

Fonte: Construção do autor

Os resultados da Tabela 5.3 indicam que o porte do estabelecimento é uma considerável fonte de variação dos preços. A variável *Nível*, diferentemente, não se mostrou relevante para justificar a variância dos preços entre estabelecimento nem diferenças nas curvas de crescimento. Entretanto, por sua importância microeconômica, ela ainda será avaliada na seção 5.3, numa tentativa de ajuste com trajetórias polinomiais.

O porte do estabelecimento, quando incorporado ao modelo, reduz em aproximadamente 20% a variância dos interceptos estimados. Ou seja, 20% da flutuação dos preços observada em outubro de 2007 pode ser explicada pela diferença de magnitude dos estabelecimento comerciais. Observa-se, também, que a variável *Porte* é capaz de descrever diferenças na evolução dos preços ao longo dos meses. Outra importante informação decorre do valor estimado para o coeficiente β_{02}, 0,64, indicando que o nível de preços em estabelecimentos de pequeno e médio porte esteve R$ 0,64 acima do estimado para os supermercados de grande porte. O critério AIC para o modelo (5.11), com a variável *Porte*, foi 1806. A comparação entre este modelo e o modelo de crescimento incondicional confirma a diferença significativa entre ambos (Tabela 5.4).

Tabela 5.4: Comparação do modelo de crescimento incondicional com o modelo com a variável "Porte".

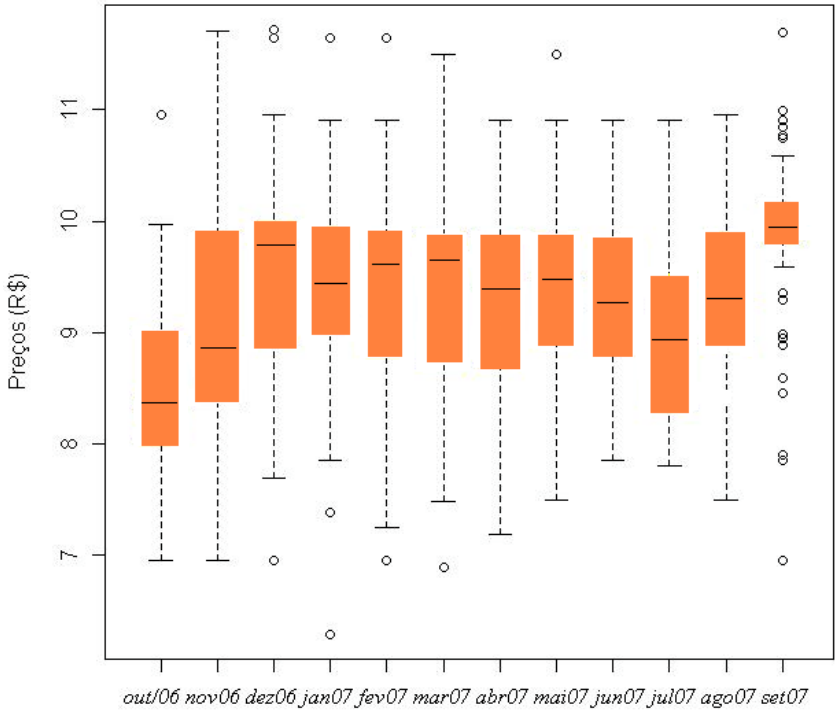
Modelo	*gl*	*AIC*	χ^2	*p-valor*
Crescimento icondicional	6	1813	12,1	0,002
Com a variável *Porte*	7	1805		

Fonte: Construção do autor

5.4 – Análise da Variância Intralocal

A observação mais atenta do *boxplot* dos preços mensais, por sua vez, remete a indícios de um comportamento não linear para a evolução dos preços ao longo dos meses (Figura 7). Por isso, além dos preditores no nível II, busca-se aprimorar o ajuste com a inclusão dos termos quadrático e cúbico da variável temporal no nível intralocal, com uma suposição, portanto, de trajetórias polinomiais para evolução dos preços.

Figura 7 – Boxplot dos preços para cada um dos doze meses pesquisados.



A especificação do modelo (5.12), abaixo, com um polinômio aleatório no nível I e com equações do nível II contendo as variáveis *Porte* e *Nível* apenas na explicação do intercepto aleatório, é capaz de gerar melhores resultados que os do modelo anterior. Sua formulação é:

$$\begin{aligned} y_{ij} &= b_{0i} + b_{1i}(t_{ij}) + \beta_2(t_{ij}^2) + \beta_3(t_{ij}^3) + \varepsilon_{ij} \\ b_{0i} &= \beta_0 + \beta_{01}P_i + \beta_{02}N + \zeta_{0i} \\ b_{1i} &= \beta_1 + \zeta_{1i} \end{aligned} \qquad (5.12)$$

$$\varepsilon_{\mathrm{ij}} = N \sim (0, \sigma_\varepsilon^2 I)$$

$$\begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix} \right)$$

onde y_{ij} é o preço do produto-local *i* no mês *j*. β_2 é o coeficiente do termo quadrático e β_3 o do termo cúbico. β_{01} e β_{02} representam, respectivamente, a influência dos preditores *Porte* e *Nível* em b_{0i} e β_1 é a média de b_{1i}. σ_0^2 é a variância de ζ_{0i} e σ_1^2 a variância de ζ_{1i}. O termo σ_{10} da matriz é a covariância entre ζ_{0i} e ζ_{1i}. ε_{ij} são erros aleatórios do nível I, para os quais se mantém a suposição de que, condicionados aos efeitos aleatórios, são independentes e normalmente distribuídos[49], com variância σ_ε^2.

Os resultados obtidos para o modelo (5.12) estão contemplados na Tabela 5.5. De imediato, percebe-se um melhor ajuste do modelo (5.12) frente ao (5.11) pelo menor valor da estatística AIC, reduzida de 1805, em (5.11), para 1712, em (5.12).

Tabela 5.5 - Resultado do ajuste do modelo 5.12 por MV.

	Parâmetro	*Estimativa*	*Desvio-padrão*	*p-valor*
Efeitos Fixos				
Intercepto	β_0	8,00	0,18	0,000
Porte (médio-baixo)	β_{01}	0,34	0,11	0,003
Nível (médio-baixo)	β_{02}	-0,27	0,13	0,043
t	β_1	0,87	0,08	0,000
t^2	β_2	-0,14	0,14	0,000
t^3	β_3	0,01	0,00	0,000
Componentes de variância			*Estimativa*	
σ_ε^2			0,39	
σ_0^2			0,37	
σ_1^2			0,0025	
Qualidade do ajuste				
AIC			1712	

Fonte: Construção do autor

[49] Esta suposição será dispensada na seção 5.6, onde uma possível correlação dos resíduos será considerada.

Outra importante mudança ocorre na estimação da variância intralocal, σ_ε^2, que se reduz de 0,47 para 0,39, com um *pseudo*-R^2 de aproximadamente 17%. É válido reparar que a condição socioeconômica da localidade do bairro onde está localizado o estabelecimento comercial, representado pela variável *Nível*, e que antes não havia sido significativa quando da estimação de (5.11), aparece, agora, com p-valor de 0,043.

O teste da razão de verossimilhança entre os modelos (5.11), com trajetória linear, e o modelo (5.12), com trajetória polinomial para a média de *y*, foi altamente significativo, com valor de 97,4 e *p-valor* inferior a $1x10^{-15}$. A tabela 5.6 ilustra.

Tabela 5.6 - Comparação das estimativas de MV dos modelos com trajetórias linear e cúbica.

Modelo	Trajetória	σ_ε^2	AIC	log-verossimilhança	χ^2
5.11	linear	0,47	1805	-894	97,4
5.12	cúbica	0,39	1712	-846	

Fonte: Construção do autor

O modelo (5.12), com média polinomial, ainda é, todavia, restritivo no que concerne à modelagem dos erros, pois a hipótese de erros independentes e com variância constante (condicionados aos efeitos aleatórios) talvez não seja a mais adequada. De fato, como já comentado em oportunidades anteriores, dados longitudinais como os provenientes de coletas mensais de preços realizadas em pesquisas de IPCs são fortes candidatos a apresentarem covariância tanto nos valores observados da variável resposta quanto nos resíduos[50].

[50] Por exemplo, Ferreira (1995, p.14) comenta que, mesmo em períodos de alta inflação, como os vividos no Brasil antes da implementação do Plano Real em 1994, grandes redes varejistas já adotavam estratégias que permitiam a permanência do preço por períodos superiores a um mês. Em períodos de inflação mais estável e em patamares mais baixos, como o caso dos anos de 2006 e 2007, que englobam o período analisado neste trabalho, é plausível a suposição de preços com valores parecidos em meses subseqüentes.

A seguir, os erros são modelados a fim de que uma eventual estrutura de covariância ainda presente possa ser contemplada no processo de estimação, possibilitando a redução da estatística AIC e, em consequencia, uma melhor qualidade do juste.

5.5 – Variância residual

É possível, portanto, que as estimativas dos resíduos apresentem variância ou correlação passíveis de serem modeladas e os modelos de efeitos mistos podem ser estendidos para este fim. Neste caso, a especificação a ser adotada é semelhante à apresentada em notação matricial em (5.4) a menos dos erros ε_{ij}, para os quais passa-se a supor uma determinada estrutura de covariância Ω_i. A notação deste novo modelo pode ser:

$$\underset{n_i x 1}{y_i} = \underset{n_i x p}{X_i}\ \underset{p \text{ x } 1}{\beta} + \underset{\text{n}_\text{i} x \text{ r}}{Z_i}\ \underset{\text{r x } 1}{\zeta_i} + \underset{\text{n}_\text{i} \text{ x } 1}{\varepsilon_i} \qquad (5.13)$$

$$\zeta_i \sim N\left(\vec{0}, \Sigma_\zeta\right)$$

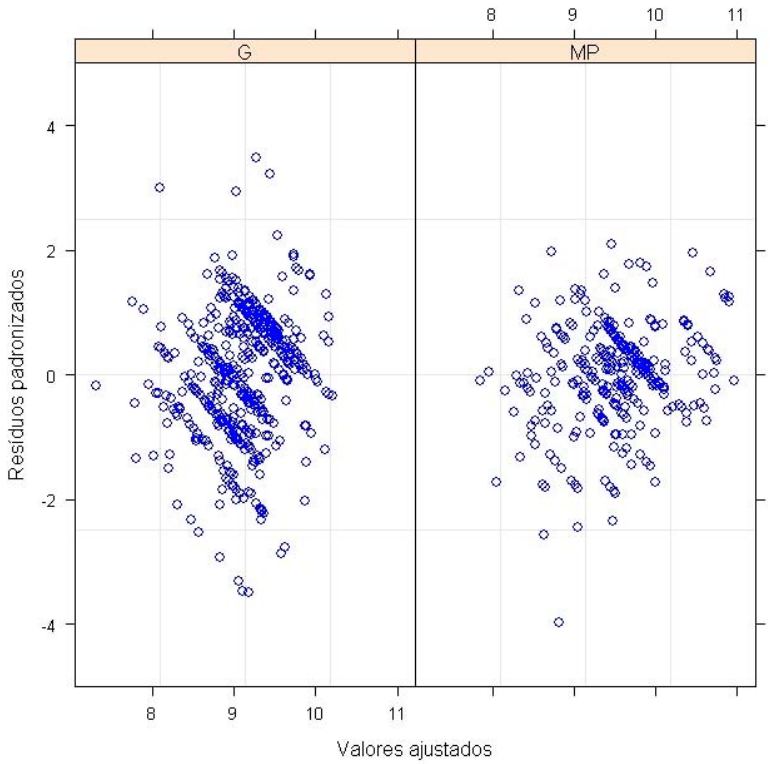
$$\varepsilon_i \sim N\left(\vec{0}, \sigma^2 \Omega_i\right)$$

A matriz Ω_i pode ser decomposta num produto de duas matrizes ($\Omega_i = V_i C_i V_i$), sendo uma V_i, relacionada à variância dos erros ao longo do tempo e, C_i, matriz de correlação com a diagonal principal formada por valores unitários[51]. Segundo Pinheiro e Bates (2000), a vantagem desta decomposição é que a variância e a correlação podem ser modeladas separadamente. A função *lme*, do R, permite este procedimento.

Iniciando com o estudo da variância, as Figuras 8 e 9 ilustram os gráficos dos resíduos padronizados do modelo (5.12) feitos separadamente para os estratos dos preditores *Porte* e

Nível. Nota-se que parece haver diferenças nas variâncias nos estratos de ambas as variáveis. A Figura 8 sugere uma amplitude ligeiramente maior nos resíduos para os preços dos estabelecimentos de grande porte se comparados aos de pequeno/médio porte.

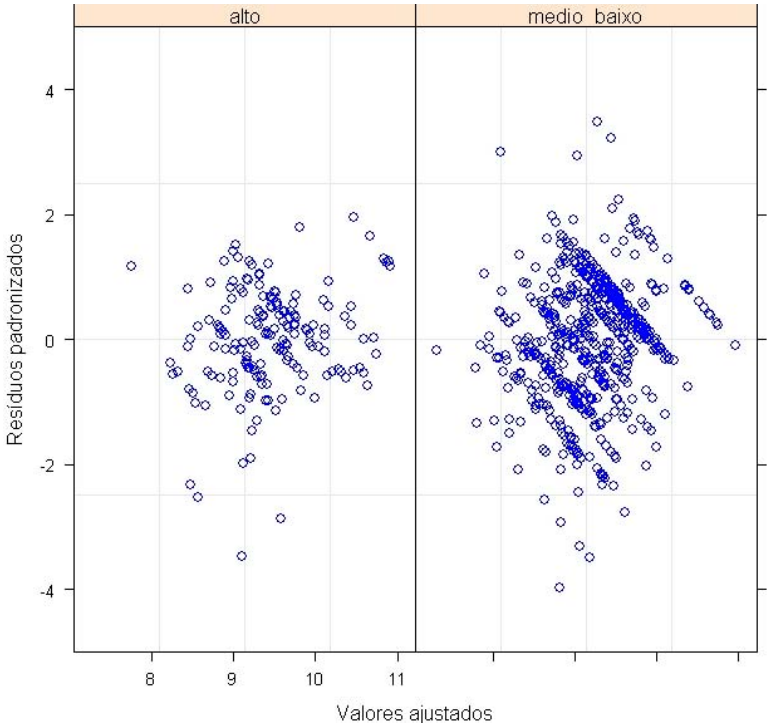
Figura 8 – Resíduos padronizados, por estrato, da variável "Porte"



Da mesma forma, o gráfico dos resíduos padronizados pelos estratos da variável *Nível* (Figura 9) mostra que a variância dos preços é aparentemente maior em localidades classificadas como sendo de nível socioeconômico medio/baixo.

[51] Pinheiro e Bates (2000, p. 205).

Figura 9 – Resíduos padronizados, por estrato, da variável "Nível"



Logo, com vista a melhoria da qualidade do ajuste, pode ser oportuna a suposição de um modelo de variância da forma[52]

$$Var(\varepsilon_{ij}) = \sigma^2 \theta^2_{(PN)ij} \ , \qquad (5.14)$$

que pode ser representado por uma função de variância do tipo $g\left((PN)_{ij}, \theta\right) = \theta_{Pij}$, onde θ é um vetor de parâmetros de variância e $\theta_{(PN)ij}$ faz alusão à variância em diferentes combinações das variáveis *Porte* (P) e *Nível* (N). A suposição da condição (5.14) à modelagem permite uma otimização da qualidade de ajuste do modelo (5.12). Para se

[52] Referências em Pinheiro e Bates (2000, p. 210)

verificar esta melhora, estima-se um modelo semelhante ao (5.12) mas que, agora, assume para os erros uma matriz $\Omega = V_i$, diagonal[53], conforme a equação abaixo:

$$\begin{aligned} y_{ij} &= b_{0i} + b_{1i}(t_{ij}) + \beta_2(t_{ij}^2) + \beta_3(t_{ij}^3) + \varepsilon_{ij} \\ b_{0i} &= \beta_0 + \beta_{01}P_i + \beta_{02}N + \zeta_{0i} \\ b_{1i} &= \beta_1 + \zeta_{1i} \end{aligned} \qquad (5.15)$$

$$\varepsilon_i \sim N\left(\vec{0}, \sigma^2 V_i\right)$$

$$\begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix}\right)$$

A Tabela 5.7 resume os resultados desta nova estimação e os compara com os do modelo (5.12). Efetivamente, se comprova a redução da estatística AIC, que cai de 1712 para 1691. $\sigma^2\Omega$

Tabela 5.7 - Comparação dos modelos 5.12 e 5.15, estimados por máxima verossimilhança.

Modelo	$\sigma^2\Omega$	*AIC*	*log-verossimilhança*	χ^2
5.12	$\sigma^2 I$	1712	-832,7	26,6
5.15	$\sigma^2 V_i$	1691	-840	

Fonte: Construção do autor

O teste da razão de verossimilhança na comparação dos dois modelos foi 26,6, relevante aos níveis usuais de significância empregados, o que confirma a melhor

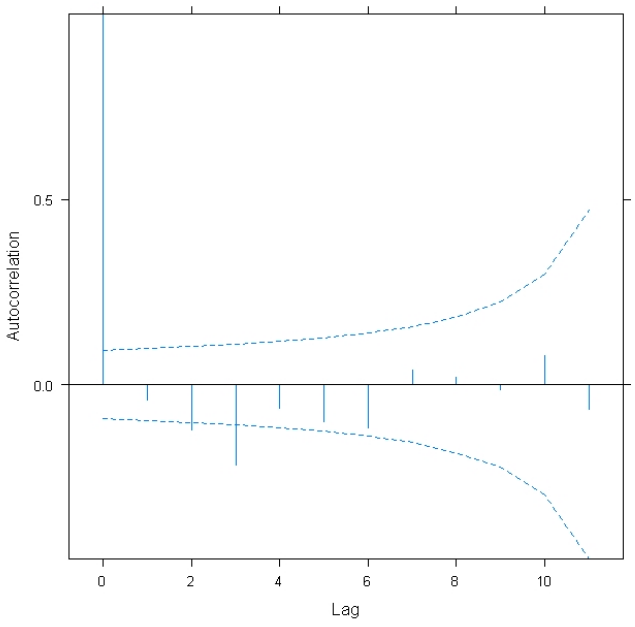
[53] Neste momento, ainda sem informações sobre as correlações, Ω é diagonal.

performance do modelo com a suposição da matriz V_i, diagonal, para os erros, ou seja, a de que a variância nos resíduos é diferente por combinações dos estratos de *Porte* e *Nível*.

5.6 – Autocorrelação

Uma vez incorporada a estrutura de variância, a atenção se volta para a modelagem da correlação que por ventura possa ainda estar presente nos resíduos. Torna-se proveitosa, então, a análise da função de autocorrelação (FAC) dos erros do modelo (5.15), apresentada na Figura 10:

Figura 10 - FAC para os resíduos do modelo 5.15:



Com a visualização da FAC, identifica-se que as autocorrelações nos lags “2” e “3”são significativas para o nível de significância de 0,01, adotado para estabelecimento dos limites inferior e superior do intervalo de confiança, representado pela linha pontilhada.

A imposição de uma matriz C_i para descrição da correlação intralocal que contemple covariâncias de segunda e terceira ordens para os erros leva, de fato, a um melhor ajuste.

A Tabela 5.8 ilustra o resultado da estatística de qualidade do ajuste, AIC, para variantes do modelo (5.15) a partir de três distintas suposições para C_i: a) identidade; b) média móvel de terceira ordem, MA(3); e c) processo autorregressivo de terceira ordem, AR(3).

Tabela 5.8 - Qualidade do ajuste para diferentes arquiteturas da matriz de correlação do modelo 5.15.

Suposição	V_i	C_i	*AIC*
a	$\sigma^2 V_i$	*I*	1691
b	$\sigma^2 V_i$	MA(3)	1683
c	$\sigma^2 V_i$	AR(3)	1684

Fonte: Construção do autor

Percebe-se que não há diferença significativa entre os dois modelos que contam com suposições mais complexas a respeito da autocorrelação. Entretanto, as reduções observadas na estatística AIC para os modelos (b) e (c) são suficientes para indicar a supremacia destes frente ao modelo mais simples (a), que supôs uma matriz diagonal para representar a estrutura de covariância dos resíduos.

Nas próximas seções, os dados são modelados a partir de suposições alternativas para a matriz de covariância Σ, numa abordagem conhecida como *modelos de padrões de covariância* (CPM) e os resultados são confrontados com os aqui obtidos.

5.7 – Modelos de Padrões de Covariância

Nos modelos de efeitos mistos até aqui apresentados, a estrutura de covariância dos *y*, $\Sigma = V(y)$, foi dada pela fórmula (5.6), construída com base em suposições a respeito dos efeitos aleatórios e, por conseguinte, da matriz Σ_ζ. Por exemplo, para um modelo aplicado a uma amostra de preços tomada em estabelecimentos comerciais durante 3 meses, com dois coeficientes aleatórios, intercepto (b_{0i}) e coeficiente angular (b_{1i}), Σ assumiria a seguinte forma:

$$\sigma_\varepsilon^2 I_{n_i} + \begin{bmatrix} \sigma_{\zeta_0}^2 & \sigma_{\zeta_0}^2 + \sigma_{\zeta_0\zeta_1} & \sigma_{\zeta_0}^2 + 2\sigma_{\zeta_0\zeta_1} \\ \sigma_{\zeta_0}^2 + \sigma_{\zeta_0\zeta_1} & \sigma_{\zeta_0}^2 + 2\sigma_{\zeta_0\zeta_1} + \sigma_{\zeta_1}^2 & \sigma_{\zeta_0}^2 + 3\sigma_{\zeta_0\zeta_1} + 2\sigma_{\zeta_1}^2 \\ \sigma_{\zeta_0}^2 + 2\sigma_{\zeta_0\zeta_1} & \sigma_{\zeta_0}^2 + 3\sigma_{\zeta_0\zeta_1} + \sigma_{\zeta_1}^2 & \sigma_{\zeta_0}^2 + 4\sigma_{\zeta_0\zeta_1} + 4\sigma_{\zeta_1}^2 \end{bmatrix}$$

Caso houvesse somente um intercepto aleatório, como em (5.8), a matriz Σ tornar-se-ia mais simples, do tipo "Compound Symmetry", com $V(y_{ij}) = \sigma_{\zeta_0}^2 + \sigma_\varepsilon^2$ e $Cov(y_{ij}, y_{ij'}) = \sigma_\zeta^2$ para $j \neq j'$ (Hedeker and Gibbons, 2006). Além dessa característica, os modelos de efeitos mistos possibilitam distinguir a variância intraindividual da variância entre estabelecimentos comerciais.

Uma outra classe de modelos, conhecida como modelos de padrões de covariância (CPM, do inglês *covariance pattern models*), onde a variável resposta também é descrita em função de covariáveis, permite o estabelecimento direto da estrutura de covariância dos erros com base em padrões conhecidos.

A representação de um modelo de padrão de covariância aplicado a um vetor resposta *y*, (n_i *x 1*), pode ser:

$$\underset{n_i\,x\,1}{y_i} = \underset{n_i\,x\,p}{X_i}\ \underset{p\,x\,1}{\beta} + \underset{n_i\,x\,1}{\varepsilon_i} \qquad (5.16)$$

$$\varepsilon_{ij} \sim N(0, \Lambda_i)$$

sendo:

$i = 1...N$: estabelecimentos;

$j = 1...n_i$: preços relacionados ao local i;

X_i : matriz de covariáveis $(n_i\ x\ p)$ do estabelecimento *i*;

β : vetor $(p\ x\ 1)$ de parâmetros;

ε_{ij} : erros normais com matriz de covariância Λ_i.

Ou seja, os y_i tem média $X_i\beta$ e sua matriz de covariância condicionada às covariáveis é Λ_i.

Jennrich and Schluchte (1986) mostraram que da mesma forma que se supõe nos modelos de coeficientes aleatórios uma estrutura de covariância para os erros, diversas possibilidades podem ser adotadas para a arquitetura da matriz Λ_i. Como exemplo desses padrões pode-se citar: modelos autorregressivos de ordem p, AR(p), Média Móvel de ordem q, MA(q), autorregressivo-média móvel, ARMA(p,q), Toeplitz (n), etc.

A Tabela 5.9 apresenta as estatísticas AIC resultantes da consideração de três diferentes padrões à matriz Λ_i do modelo 5.16, quais sejam, AR(1), MA(1) e Não Estruturada (NE), além do resultado do modelo de efeitos aleatórios com estrutura de covariância do tipo AR(3) (item c da Tabela 5.8), considerado como sendo o mais satisfatório no leque dos

estimados até aqui. A Tabela 5.9 é complementada com as características dos quatro modelos.

Tabela 5.9 - Comparação dos modelos CPM com o modelo de efeitos aleatórios 5.15c.

Modelo	Σ				
	Σ_ζ	Ω	V_i	C_i	*AIC*
5.15 (c)	Efeitos aleatórios	$\sigma^2 V_i C_i V_i$	$\theta^2((PN)_{ij})$	*AR(3)*	1684
5.16 (a)	____	$AR1 + \sigma^2 V_i$	$\theta^2((PN)_{ij})$	*AR(1)*	1878
5.16 (b)	____	$MA1 + \sigma^2 V_i$	$\theta^2((PN)_{ij})$	*MA(1)*	1879
5.16 (c)	____	Não Estruturada	Não Estruturada	Não Estruturada	1707

Fonte: Construção do autor

É visível a melhor qualidade do ajuste do modelo com efeitos aleatórios para a amostra considerada. A estimação deste modelo apresentou AIC mais baixo mesmo que o do modelo (5.16c), com matriz de covariância não estruturada.

A opção pelo modelo (5.15c) parece ainda mais acertada se a análise recair sobre a parcimônia dos dois modelos. A estimação do modelo com matriz não estruturada consumiu 76 graus de liberdade, enquanto que o modelo (5.15c) apenas 16.

No próximo capítulo, será feita uma avaliação sobre o uso dos modelos longitudinais para fins de imputação de não-resposta em IPC.

Capítulo 6 – Imputação

No terceiro capítulo, foram descritos aspectos conceituais relacionados à não-resposta e aos cuidados que analistas devem ter quando lidam com conjuntos de dados incompletos. Como foi lá destacado, um dos principais problemas é a impossibilidade de aplicação de alguns métodos analíticos quando não se pode supor que a não-resposta seja completamente aleatória (MCAR)[54]. Na situação aqui estudada, a hipótese é a de que a não-resposta é dependente de pelo menos um dos preditores, como o "Porte" dos estabelecimentos.

Em dados de pesquisas de índices de preços ao consumidor, assim como em diversos estudos longitudinais, a não-resposta é um fenômeno comum. No SNIPC, por exemplo, altas taxas de preços omissos são registradas todos os meses. No caso da amostra do subitem arroz, utilizada como base para realização deste trabalho, as taxas de não-resposta para algumas marcas pesquisadas no período de outubro de 2006 a setembro de 2007 ultrapassaram 50%. Adicionalmente, foram verificadas importantes diferenças nas taxas de não-resposta em função do porte do estabelecimento comercial. Para o produto de código *05000*, analisado nos modelos estimados no capítulo anterior, a taxa de não-resposta global foi de 11,4%, sendo que, no estrato de grandes supermercados, o percentual foi de 9,3 e, no de estabelecimentos de porte pequeno ou médio, de 15,1. Uma regressão logística aplicada aos dados confirmou a significância das diferenças entre as probabilidades de não-resposta[55] segundo cada uma das cinco marcas avaliadas.

Estes números sugerem, fortemente, que os procedimentos de imputação estão sempre muito presentes nas pesquisas de IPC e que seus impactos, portanto, não devem ser desprezados.

[54] Ver capítulo 3 para comentários sobre os tipos de não-resposta.
[55] As estimativas da regressão logística estão no Anexo 2.

Com o intuito de se mostrar como os modelos longitudinais podem ser empregados para imputação de não-resposta, inclusive por poderem ser usados mesmo em situações onde outras técnicas não são aplicáveis, promove-se, neste capítulo, a avaliação de erros quadráticos médios e de medidas de vício[56] resultantes das comparações entre dados reais da amostra e valores preditos pelo modelo (5.15c), que considera que o porte do estabelecimento e a localização da empresa varejista afetam os níveis de preços.

Os resultados são também comparados com os gerados por três outras formas de imputação normalmente adotados em pesquisas de índices de preços ao consumidor e que são: a) imputação do preço do bem i no mês t pela média dos preços observados em t; b) imputação pelo último preço observado; e c) imputação pela variação média dos preços observados no mês t.

Para comparação dos diferentes métodos, realizou-se uma validação cruzada que consistiu na seleção de dez amostras independentes com probabilidades diferenciadas em função do porte dos estabelecimentos e com base em uma suposição de não-resposta do tipo aleatória (MAR).

Em caráter complementar, foram consideradas outras combinações para as probabilidades de não-resposta do produto *05000* em função do porte dos estabelecimentos. Neste cenário, foram estimadas, para cada método de imputação e para cada uma das dez amostras selecionadas, medidas de vícios e de erros quadráticos médios.

Nas seções 6.1 a 6.4, são apresentados os quatro procedimentos estudados considerando-se probabilidades de perda para estabelecimentos de grande porte (ϕ_g) e para

[56] Foram calculados vícios absolutos $\left(\sum_{i=1}^{n_s}\frac{\hat{p}_i - p_i}{p_i}\right)$ e relativos $\left[\left(\sum_{i=1}^{n_s}\frac{|\hat{p}_i - p_i|}{p_i}\right)\Big/ n_s\right]$ para cada amostra s, s=1:10.

estabelecimento de porte pequeno ou médio (ϕ_{mp}) como sendo próximas às taxas encontradas no período, respectivamente, de 10% e de 15%[57].

6.1 – Imputação pela média

Quando se aplica a imputação pela média, está se admitindo que o movimento do preço do produto ausente é o mesmo que o da média dos demais itens que compõem a amostra. Vista de outra forma, a imputação pela média se enquadra dentro do que na literatura se chama de análise de casos completos[58]. Neste procedimento, toda informação para realização da imputação é "transversal", descartando qualquer conhecimento da relação entre o preço do produto cujo preço se deseja imputar e o preço dos demais em momentos anteriores. Em outras palavras, admite-se a forte suposição de que há exata correlação entre o preço omitido e a média dos demais. Se esta hipótese não for verdadeira, e quase que invariavelmente não é em estudos longitudinais, então a imputação pela média é equivocada. Demirtas (2004) é enfático:

> Imputing the subject-mean seriously distorts trends over time and within-subject covariance structure. Imputing the occasion-mean distorts trends within subjects and between subject variation. Both mean imputation methods introduce bias into longitudinal analyses and seriously impair standard errors and hypothesis tests. (DEMIRTAS, 2004, p. 308).

No domínio da análise de preços, a citação de Demirtas (2004) leva à percepção de como a imputação pelo preço médio do produto pode distorcer a estrutura de covariância intralocal e prejudicar testes de hipóteses. Outra consequencia da imputação pela média ainda

[57] As taxas verificadas foram de 9,3% para estabelecimentos de grande porte e de 14,7 para de pequeno/médio porte.
[58] Procedimento conhecido como complete case analysis, Little e Rubin (1987); Diggle, Liang e Zeger (1994); Molenberghs et al. (2004).

mais evidente é que a omissão do preço do bem *i* no mês *t* conduz a um aumento dos pesos (implícitos) dos produtos cotados.

Matematicamente, o preço do bem *i*, quando imputado pela média, no tempo *t*, $\hat{p}_i^t$, é dado por:

$$\hat{p}_i^t = \frac{\sum_{j=1}^{n} p_j^t}{n} \qquad i \neq j \qquad (6.1)$$

A aplicação da imputação pela média e a comparação com os valores observados para as 10 amostras, que foram selecionadas considerando-se uma probabilidade de perda de 10% para estabelecimentos de grande porte (ϕ_g = 0,10) e de 15% para estabelecimentos pequenos ou médios (ϕ_{mp} = 0,15), levou aos resultados de erros quadráticos médios e de vícios ilustrados pela Tabela (6.1).

Tabela 6.1 - EQMs, preços médios imputados, preços médios observados e estimativas de vício para imputação pela média. (ϕ_g = 0,10; ϕ_{mp} = 0,15)

Amostra (s)	*EQM (s)*	*média dos preços imputados*	*média dos preços observados*	*dif (vício abs)*	*Vício relativo*
1	0,69	9,2	9,30	-0,10	0,07
2	0,50	9,22	9,32	-0,10	0,06
3	0,86	9,22	9,33	-0,10	0,08
4	0,55	9,24	9,28	-0,03	0,06
5	0,73	9,24	9,32	-0,07	0,07
6	0,84	9,23	9,19	0,03	0,08
7	0,69	9,25	9,51	-0,26	0,07
8	0,56	9,22	9,30	-0,07	0,07
9	0,57	9,26	9,34	-0,07	0,07
10	0,91	9,29	9,13	0,16	0,09
Geral	0,67	9,24	9,30	-0,06	0,07

Fonte: Construção do autor

A última linha da Tabela 6.1 apresenta os resultados globais calculados. Na quinta coluna, merece atenção a diferença entre os preços médios imputados e observados, de -0,06. Ou seja, a adoção deste procedimento, no caso da amostra estudada, e para as suposições adotadas, levaria a valores imputados cerca de seis centavos, em média, abaixo dos valores reais. O EQM da imputação pela média, com probabilidade de perda nos estabelecimentos de porte pequeno/médio 50% acima da dos de grande porte, foi de 0,67.

Apesar de ser um método fácil de ser aplicado e interpretado, a imputação pela média apresenta o sério problema de gerar estimativas viciadas em algumas situações, principalmente quando a não-resposta é do tipo MAR, ou seja, quando não se pode supor independência entre a ausência do preço e alguns preditores. Se, por exemplo, a taxa de não-resposta for maior em estratos formados por pequenos estabelecimentos do que em estratos compostos por estabelecimentos de grande porte, como a aqui observada, a imputação pela média não deve ser adotada[59].

6.2 – Imputação pelo último preço observado

O valor imputado pelo método do último preço observado para um bem *i* no mês *t*, $\hat{p}_i^t$, pode ser representado, em termos formais, por:

$$\hat{p}_i^t = p_i^{t-j} \qquad (6.2)$$

onde p_i^{t-j} é o preço observado do bem i no mês *t-j* , tal que *j* corresponde ao número de meses entre o mês *t* e o último mês em que o preço do produto *i* foi coletado, inclusive.

[59] Para referências, ver capítulo 3.

Os erros quadráticos médios e estimativas de vício recorrendo-se a este procedimento, para imputação nas 10 amostras selecionadas, bem como os preços médios, estão na Tabela 6.2. Verifica-se que o erro quadrático médio, 0,62, foi menor do que o EQM da imputação pela média. Já o vício, em termos absolutos (-0,08), foi maior. Ambos os métodos tendem a subestimar os valores verdadeiros.

A imputação pelo último preço observado, ainda que adotada na prática internacional de índices de preços[60], é também um procedimento comprometedor da qualidade dos resultados dos índices elementares, devendo ser evitado. Segundo ILO et al (2004, p. 160), "carrying forward the last observed price should be avoided wherever possible". Diante de períodos de alta inflação ou de rápidas alterações de preços no mercado, esta abordagem é ainda mais problemática. Um ponto a ser destacado sobre a imputação pelo último preço observado é que sua prática leva, normalmente, a subestimativas ou, em casos limites, a resultados nulos dos indicadores.

Tabela 6.2 - EQMs, preços médios imputados, preços médios observados e estimativas de vício para imputação pelo último preço observado. ($\phi_g = 0{,}10; \phi_{mp} = 0{,}15$)

Amostra (s)	*EQM (s)*	*média dos preços imputados*	*média dos preços observados*	*dif (vício abs)*	*Vício relativo*
1	0,56	9,17	9,30	-0,13	0,05
2	0,53	9,11	9,32	-0,20	0,05
3	0,43	9,32	9,33	0,00	0,04
4	0,58	9,22	9,28	-0,05	0,04
5	0,55	9,21	9,32	-0,10	0,05
6	0,61	9,13	9,19	-0,06	0,05
7	0,85	9,44	9,51	-0,07	0,06
8	0,52	9,09	9,30	-0,20	0,04
9	0,76	9,25	9,34	-0,09	0,06
10	0,77	9,25	9,13	0,11	0,06
Geral	0,62	9,22	9,30	-0,08	0,06

Fonte: Construção do autor

[60] ILO et al. (2004).

6.3 – Imputação pela variação média

Imputar pela variação média do agregado elementar é multiplicar o preço do bem i no mês t-1 pela variação média observada no mês t dos produtos pesquisados. É um procedimento que pode ser aplicado quando se espera que os preços se movam na mesma direção e equivale à omissão do item do cálculo efetuado. De acordo com o manual de índices de preços ILO et al. (2004), esta forma de imputação é mais adequada do que as duas anteriores. Este método também é de fácil implementação. Analiticamente, tem-se que o preço estimado do bem *i* em *t*, $\hat{p}_i^t$, é:

$$\hat{p}_i^t = p_i^{t-1}.R_\Theta^t$$

onde p_i^{t-1} é o preço do bem i no mês anterior e R_Θ^t é a variação média dos preços dos produtos do conjunto Θ, formado por aqueles cujos preços foram cotados em t e em t-1[61].

Como $\hat{p}_i^t$ depende de p_i^{t-1}, é importante que a imputação sempre seja feita, a fim de que a base de preços esteja disponível para eventuais imputações subsequentes.

A Tabela 6.3 contém os resultados da validação cruzada aplicada para os valores estimados por este método. É interessante observar que a estimativa do vício absoluto (0,04) com a adoção desta metodologia, teve valor mais baixo que os métodos anteriormente apresentados. Entretanto, o erro quadrático médio (0,75), muito influenciado pelo resultado da décima amostra, foi maior do que o calculado para a imputação pela média e também superior ao do método que imputa pelo último valor observado.

[61] Sendo que a base de dados do mês anterior pode conter preços imputados pelo mesmo método.

Tabela 6.3 - EQMs, preços médios imputados, preços médios observados e estimativas de vício para imputação pela variação média. ($\phi_g = 0{,}10; \phi_{mp} = 0{,}15$)

Amostra (s)	*EQM (s)*	*média dos preços imputados*	*média dos preços observados*	*dif (vício abs)*	*Vício relativo*
1	0,70	9,3	9,30	0,00	0,07
2	0,52	9,34	9,32	0,02	0,06
3	0,92	9,33	9,33	0,00	0,08
4	0,63	9,32	9,28	0,04	0,06
5	0,69	9,35	9,32	0,04	0,07
6	0,95	9,29	9,19	0,09	0,08
7	0,72	9,36	9,51	-0,15	0,07
8	0,72	9,38	9,30	0,08	0,07
9	0,65	9,39	9,34	0,05	0,07
10	1,03	9,36	9,13	0,23	0,09
Geral	0,75	9,34	9,30	0,04	0,07

Fonte: Construção do autor

6.4 – Imputação pelo Modelo Longitudinal

Uma das aplicações oferecidas pela modelagem longitudinal, como descrito na introdução, é a possibilidade de uso dos valores preditos para imputação. Os modelos de efeitos mistos, como os estimados no quinto capítulo, são ainda valiosos pelo fato de serem robustos em situações onde outros métodos falham. Por exemplo, podem mesmo ser usados em situações onde a não-resposta é MAR e quando o conjunto de dados é não-balanceado.

Na Tabela 6.4, estão os erros quadráticos médios e medidas de vício estimadas na comparação dos valores preditos pelo modelo ajustado (5.15c)[62] com os valores observados. O EQM geral, 0,45, foi bem menor do que os demais e a estimativa do vício absoluto teve valor bem próximo de zero (0,002). A média geral dos preços imputados coincidiu, até a segunda casa decimal, com a média geral dos preços observados.

[62] A metodologia para obtenção de estimativas de máxima verossimilhança para os coeficientes individuais, *BLUPs* (Best Linear Unbiased Predictors), é encontrada em Pinheiro e Bates (2000).

Tabela 6.4 - EQMs, preços médios imputados, preços médios observados e estimativas de vício para imputação pelo modelo longitudinal. $(\phi_g = 0{,}10; \phi_{mp} = 0{,}15)$

Amostra (s)	*EQM (s)*	*média dos preços imputados*	*média dos preços observados*	*dif (vício abs)*	*Vício relativo*
1	0,40	9,27	9,30	-0,03	0,05
2	0,38	9,24	9,32	-0,07	0,05
3	0,49	9,38	9,33	0,05	0,05
4	0,41	9,34	9,28	0,06	0,05
5	0,54	9,34	9,32	0,02	0,06
6	0,55	9,18	9,19	0,00	0,06
7	0,39	9,42	9,51	-0,08	0,05
8	0,40	9,23	9,30	-0,06	0,05
9	0,36	9,31	9,34	-0,03	0,05
10	0,54	9,3	9,13	0,17	0,06
Geral	0,45	9,30	9,30	0,00	0,06

Fonte: Construção do autor

A Figura 11 ilustra o gráfico com os valores de vício para cada uma das 10 amostras e para cada um dos quatro métodos quando a probabilidade de não-resposta nos estabelecimentos pequenos ou médios supera em 50% a dos de grande porte. A Tabela 6.5 apresenta um resumo dos resultados.

Como destacado na Tabela 6.5, o EQM estimado com imputações pelo modelo longitudinal foi bem inferior aos demais. Além disso, as medidas calculadas de vício para o modelo e para o método da variação média, que também leva em conta informações longitudinais e transversais, foram melhores que as outras duas. Observando-se a Figura 11, nota-se que a flutuação das estimativas de vício em cada uma das amostras para o modelo longitudinal é próxima da nulidade. As imputações pela média e pelo último preço observado, além de altos EQMs, apresentaram vício sistemático.

Figura 11 - Estimativas das diferenças entre os preços médios estimados e observados, segundo cada um dos quatro métodos.

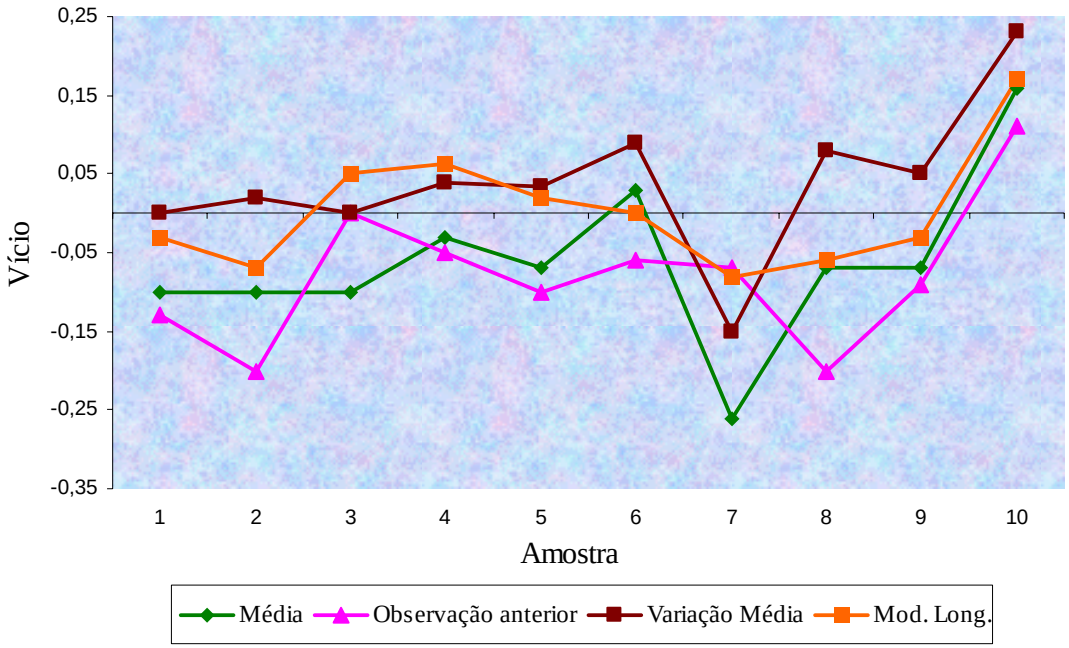


Tabela 6.5 - EQMs gerados pela adoção de quatro métodos de imputação com Pg(0,10) e Ppm(0,15) $(\phi_g = 0{,}10; \phi_{mp} = 0{,}15)$

Estatística	*Modelo longitudinal*	*Média*	*Último preço observado*	*Variação média*
EQM	0,45	0,69	0,62	0,75
Vício	0,00	-0,06	-0,08	0,04

Fonte: Construção do autor

6.5 – Estimativas de vício para outras relações entre as probabilidades de não-resposta

Nas seções anteriores, as dez amostras selecionadas para validação cruzada foram obtidas levando-se em consideração probabilidades diferenciadas para não-resposta em função do tipo de estabelecimento. Foi estipulada probabilidade de perda de 0,15 para supermercados de porte pequeno ou médio e de 0,10 para os de grande porte (diferença de 50%), com $\phi_{mp}/\phi_g = 1{,}50$. Estes valores foram determinados a partir de taxas de não-resposta observadas.

Entretanto, dois comentários cabem aqui: primeiro, as taxas calculadas do produto *05000* não são das mais altas observadas. Como mencionado na introdução deste capítulo, outros produtos apresentaram taxas de não-resposta muito maiores; segundo, essas taxas podem mudar em decorrência da dinâmica do mercado, sujeito à influência de uma série de fatores, de modo que a razão entre as probabilidades de perda atribuídas a diferentes tipos de estabelecimentos, ϕ_{mp}/ϕ_g, pode ter alta variância.

Nesta seção, são apresentados os resultados das estatísticas de vício para os quatro métodos analisados de acordo com dez combinações de probabilidades de não-resposta segundo o tipo de estabelecimento. Mais especificamente, são feitas simulações para razões ϕ_{mp}/ϕ_g iguais a 1,0, 1,1, 1,2, 1,3, 1,4, 1,5, 1,6, 1,7, 1,8, 1,9 e 2,0. O valor 1 para a razão ϕ_{mp}/ϕ_g =1 indica que se supõe que a não-resposta é independente do porte do estabelecimento. A Tabela 6.6 mostra as estimativas de vício para cada um dos quatro métodos de imputação segundo os valores assumidos para a razão ϕ_{mp}/ϕ_g.

Tabela 6.6 - Estimativas de vício em função das razões ϕ_{mp} / ϕ_g

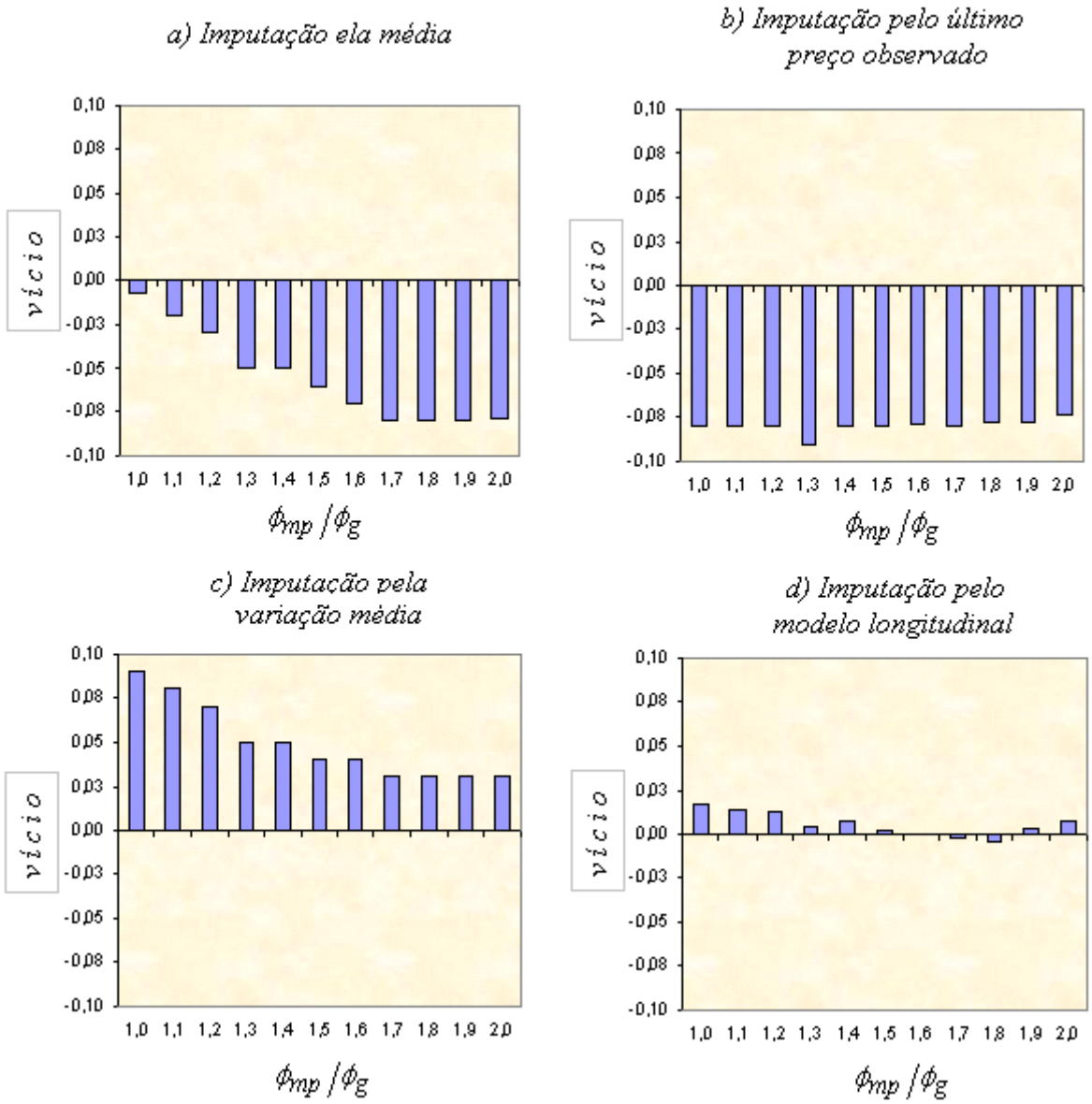
ϕ_{mp} / ϕ_g	*Média*	*Último preço observado*	*Variação média*	*Modelo*
1,0	-0,01	-0,08	0,09	0,02
1,1	-0,02	-0,08	0,08	0,01
1,2	-0,03	-0,08	0,07	0,01
1,3	-0,05	-0,09	0,05	0,00
1,4	-0,05	-0,08	0,05	0,01
1,5	-0,06	-0,08	0,04	0,00
1,6	-0,07	-0,08	0,04	0,00
1,7	-0,08	-0,08	0,03	0,00
1,8	-0,08	-0,08	0,03	0,00
1,9	-0,08	-0,08	0,03	0,00
2,0	-0,08	-0,07	0,03	0,01

Fonte: Construção do autor

Sobressaem, na Tabela 6.6, os reduzidos valores das estimativas de vício do modelo longitudinal. (coluna “Modelo”). Os valores estimados quando se adota o modelo, que leva em conta o porte do estabelecimento na predição, são, para quase todos os valores da razão ϕ_{mp} / ϕ_g , bem próximos de zero. Ao contrário, os outros métodos apresentam resultados insatisfatórios. A imputação pela média, como se nota na segunda coluna da Tabela, aumenta o grau de subestimação quando a relação ϕ_{mp} / ϕ_g aumenta, ou seja, quanto menos provável é a hipótese de não-resposta do tipo MCAR. O método de imputação pelo último preço observado apresentou fortes resultados negativos, com estimativas de vício em torno de -0,08 para quase todas as combinações de ϕ_{mp} / ϕ_g . Já a imputação pela variação média (coluna 4), mostrou uma ligeira tendência de superestimação dos resultados, porém, com a magnitude do vício menor do que as apresentadas pelos métodos da imputação pela média e pela última observação. As Figuras 12-a, 12-b, 12-c e 12-d, a seguir, ilustram os resultados de vício para

os quatro métodos em função da razão ϕ_{mp}/ϕ_g. A visualização dos gráficos facilita a percepção do melhor desempenho do modelo longitudinal frente aos demais métodos.

Figura 12 (a, b, c e d) - Vícios estimados dos diferentes métodos de imputação em função da razão ϕ_{mp}/ϕ_g



Fonte: Construção do autor

O estudo aqui realizado de comparação entre quatro abordagens passíveis de uso em pesquisas de IPC[63] teve, por propósito, ilustrar mais uma possibilidade de aplicação dos modelos longitudinais em índices de preços. Contudo, deve-se lembrar a limitação do trabalho, restrito ao agregado elementar, e da avaliação empírica, baseada na amostra de apenas um subitem. Não obstante, os resultados encontrados estão de acordo com o esperado teoricamente.

No próximo capítulo, estende-se a aplicação dos modelos longitudinais para uma amostra mais heterogênea, composta por cinco marcas do subitem *arroz*. Mostra-se como os índices de Jevons e de Dutot, calculados a partir de estimativas mais depuradas, controladas pela heterogeneidade, são mais próximos entre si do que quando calculados de forma tradicional, reforçando a teoria de que a dispersão dos preços, que pode ser modelada, como de fato tem-se feito aqui, é uma importante fonte de diferenças nos resultados das principais fórmulas empregadas no cálculo do nível mais desagregado em IPCs.

[63] Foi também avaliada a imputação pelo algoritmo EM (DEMPSTER, LAIRD E RUBIN, 1977), com resultados insatisfatórios. Esta simulação não foi mostrada aqui por sua aplicação fugir ao contexto dos IPCs.

Capítulo 7 – Índice de Dutot com Heterogeneidade Controlada por Modelos Hedônicos

Os índices de Dutot e de Jevons são os mais usados por institutos oficiais de estatística para cálculo no agregado elementar. Esses índices normalmente geram, se aplicados a agregados homogêneos, resultados muito próximos. Entretanto, uma das diferenças entre eles é que o índice de Dutot falha no teste da comensurabilidade, superestimando, via de regra, as estimativas obtidas[64].

No segundo capítulo, foram descritas as principais características de ambos os índices, com ênfase para as visões axiomática e estocástica desses indicadores. Lá, foi também apresentada a equação que descreve a relação aproximada entre eles, por aproximação de Taylor[65]:

$$I_J(p^{t-1}, p^t) \approx I_D(p^{t-1}, p^t)[1 + (1/2)\,\mathrm{var}(e^{t-1}) - (1/2)\,\mathrm{var}(e^t)] \qquad (7.1)$$

onde I_J e I_D são os índices de Jevons e de Dutot e $\mathrm{var}(e^{t-1})$ e $\mathrm{var}(e^t)$ representam, respectivamente, as variâncias dos desvios de preços (p^{t-1} e p^t) em torno da média nos períodos (t-1) e (t).

Como se nota na expressão (7.1), são as variâncias dos preços nos meses t e t-1 que determinam o quão diferentes são os resultados dos índices. Com a redução destas variâncias, espera-se uma aproximação entre eles. Neste capítulo, propõe-se a utilização das variâncias intralocais nos meses t e $t-1$, (ξ_t^2 e ξ_{t-1}^2), estimadas por *modelos hedônicos com coeficientes aleatórios,* líquida da influência dos estabelecimentos, para cálculo de índices com heterogeneidade (dispersão) controlada. Espera-se que os índices de Jevons e de Dutot,

[64] Ver seção 2.2.2.
[65] Consultar Diewert (1995) ou Dalen (1992) para referências.

estimados a partir dessa regressão hedônica, sejam mais próximos entre si do que os tradicionalmente calculados.

7.1 – Modelagem de um Agregado Elementar Mais Heterogêneo

No capítulo anterior, onde os modelos foram estimados com o intuito de se mostrar que valores por eles preditos podem ser usados para imputação sob a condição de não-resposta do tipo MAR, a amostra foi composta por produtos de uma só marca. Uma das justificativas ali apresentadas foi a de que alguns outros procedimentos de imputação, com os quais foram comparados os resultados da imputação gerada pelo modelo longitudinal, aplicam-se a estratos homogêneos, formados, em muitas situações, por apenas uma marca de produtos.

Neste sétimo capítulo, sendo o objetivo a avaliação da relação entre os índices de Jevons e de Dutot, estende-se a amostra para um agregado elementar composto por diferentes marcas, mais precisamente, cinco marcas do subitem arroz.[66]

Portanto, esta etapa do trabalho se desenvolve da seguinte forma: primeiro, identificam-se, por intermédio da modelagem longitudinal, preditores importantes na análise dos preços; em seguida, aplicam-se modelos hedônicos hierárquicos para períodos de dois meses com o objetivo mais específico de se extrair a variação pura de preços.

7.1.1 – Preditores

Com a ampliação da amostra para cinco marcas, três foram os fatores cujas significâncias foram analisadas na predição dos níveis de preços: o porte do estabelecimento

[66] A fórmula de Jevons, ao contrário da de Dutot, não exige agregados homogêneos.

onde o produto é pesquisado[67] (variável *P*), a classificação socioeconômica do bairro onde se localiza (variável *N*) e a marca do produto (variável *M*).

Após a estimação de alguns modelos, percebeu-se que a localização do estabelecimento comercial não é fator preponderante para determinação dos níveis de preços nem das variâncias entre estabelecimentos quando se tem por base a amostra estudada. Assim, foram identificados como relevantes para análise o porte do estabelecimento e a marca do produto, sendo que a marca influenciou, ainda, a evolução dos preços nos meses considerados.

O modelo estimado (7.2), cuja equação está reproduzida a seguir, contemplou coeficientes aleatórios (b_{0i} e b_{1i}), estrutura de correlação MA(3) e função de variância do tipo $g(w_{ij},\theta) = |\, w_{ij} \,|^{\theta}$, onde os w_{ij} foram estimados através de um indicador de disponibilidade da marca no mês *j*, calculado pelo percentual de locais em que efetivamente o preço do produto *i* foi coletado em relação ao total de locais que o comercializavam no referido mês[68].

$$\begin{aligned} y_{ij} &= b_{0i} + b_{1i}(t_{ij}) + \beta_2(t_{ij}^2) + \beta_3(t_{ij}^3) + \varepsilon_{ij} \\ b_{0i} &= \beta_0 + \beta_{01}P_i + \beta_{02}M_i + \zeta_{0i} \\ b_{1i} &= \beta_1 + \beta_{12}M_i + \zeta_{1i} \end{aligned} \qquad (7.2)$$

Em (7.2), y_{ij} é o preço do produto-local *i* no mês *j*. β_2 é o coeficiente do termo quadrático e β_3 o do termo cúbico. β_{01} e β_{02} representam, respectivamente, a influência dos

[67] Relembra-se, a variável usada como *proxy* do porte foi "*Pessoal Ocupado - PO"*, do Cadastro Central de Empresas (CEMPRE) do IBGE.

[68] A teoria econômica suporta uma variabilidade dos preços de produtos como função da sua disponibilidade no mercado recorrendo a uma teoria não explorada aqui em mais detalhes, a teoria do custo de busca (Silver e Heravi (2003; 2007); Varian(1993). Uma das suposições subjacentes à teoria do custo de busca é a de que a menor presença no mercado de um bem, representada por poucos vendedores ou pouca oferta, aumenta seu custo de busca e abre espaço para que a firma possa usufruir desta situação, aproveitando para aumentar preços na tentativa de alcançar maiores lucros. No caso de se transportar a idéia para a situação particular aqui estudada, uma marca muito presente, com um alto w_{ij}, pode ter variância menor do que outra ofertada, por exemplo, a um

preditores *Porte* e *Marca* em b_{0i}. $\beta_{12,}$ com quatro níveis, é o efeito da marca no coeficiente aleatório b_{1i}. σ_0^2 é a variância de ζ_{0i} e σ_1^2 a variância de ζ_{1i}. O termo σ_{10} da matriz é a covariância entre ζ_{0i} e ζ_{1i}. Os erros e os efeitos aleatórios são normalmente distribuídos, com as seguintes características:

$$\varepsilon_{ij} \sim N\left(\vec{0}, \sigma^2 \Omega_i\right)$$

$$\begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix}\right)$$

A Tabela 7.1 contém os principais resultados deste modelo.

Tabela 7.1 - Algumas das estimativas do modelo 7.2 por MV.

Coeficiente	*Efeitos fixos*		*Componentes de variância*				
	Estimativa	*p-valor*					
Intercepto	7,03	0,000		*Notação*	*Estimativa*	*Limite inferior*	*Limite superior*
Mês	0,78	0,000	Entre locais	σ_0^2	0,2700	0,18	0,38
Porte (médio-baixo)	0,12	0,075		σ_1^2	0,0013	0,0005	0,003
Marca "05000"	0,89	0,000	Intralocal	σ_ε^2	0,4400	0,38	0,5
Marca "07000"	-0,53	0,002					
Marca "09000"	0,30	0,041					
Marca "10000"	0,09	0,547					

Fonte: Construção do autor

Verifica-se, com a Tabela (7.1), que os preços dos produtos em estabelecimentos de porte pequeno ou médio são, em média, R$ 0,12 (doze centavos) acima dos preços praticados em supermercados de grande porte. Além disso, os níveis de preços das cinco marcas

particular grupo da população ou pouco comercializada. Decorre daí a modelagem da variância de acordo com a função *g* proposta.

analisadas também diferem de forma considerável, com exceção da marca de código “10000”, que não apresentou discrepância significativa em relação à marca de comparação, “02000”, com *p-valor* de 0,547. Em relação às componentes de variância, a intralocal residual (0,44) é maior que a observada entre locais (0,27). A variância não explicada em b_{1i} é 0,0013. Entretanto, todas são ainda estatisticamente diferentes de zero.

7.2 – Estimação do IPC por regressão hedônica

Os preditores determinados na análise feita na seção anterior (*Porte e Marca*) serão incluídos, aqui, em modelos hedônicos para estimação dos índices de Jevons e de Dutot com dispersão controlada tanto pela heterogeneidade de produtos quanto de supermercados.

No enfoque estocástico não ponderado de IPC[69], o logaritmo da razão de preços entre os períodos t e t-1 pode ser dado por:

$$\ln(\frac{p_i^t}{p_i^{t-1}}) = \beta + \varepsilon_{i;} \qquad i = 1,2,...,n \qquad (7.3)$$

onde o coeficiente β *é* um estimador não viciado do logaritmo da taxa de inflação e ε_i são variáveis aleatórias independentes com média 0 e variância constante σ^2. Pode-se demonstrar, também, que o estimador de máxima verossimilhança de β é o logaritmo da média geométrica dos preços relativos (ILO et al., 2004, p. 300). Logo, exp($\hat{\beta}$) é equivalente ao índice de Jevons entre t e t-1, ou, matematicamente:

[69] Edgeworth (1923), Diewert (1993).

$$\exp(\hat{\beta}) = I_J\left(\frac{p^t}{p^{t-1}}\right) = \prod_{i=1}^{n} \sqrt[n]{\frac{p_i^t}{p_i^{t-1}}} \qquad (7.4)$$

Se, da mesma forma que o logaritmo da razão de preços em (7.3), forem modelados por intermédio de uma regressão hedônica com efeitos aleatórios os logaritmos dos preços do mês τ (τ = t-1, t), mas agora levando-se em conta os preditores *Porte (P)* e *Marca (M),* pode-se estabelecer a seguinte equação hedônica:

$$\begin{aligned} \ln(p_i^t) &= b_{0i} + \beta_1 D^t + \varepsilon_i^t \\ b_{0i} &= \beta_0 + \beta_{01} P_i + \beta_{02} M_i + \zeta_0 \end{aligned} \qquad (7.5)$$

com b_{0i} representando interceptos aleatórios, β_1 o coeficiente da variável dicotômica D^t, que assume valor 1 se o mês é t e 0 se o mês é t-1, e β_{01} e β_{02} as influências em b_{0i} dos preditores *Porte* e *Marca*, respectivamente. Os erros ε_i^t são normalmente distribuídos com média 0 e variância σ_ε^2, enquanto que os efeitos aleatórios ζ_{0i} tem média 0 e variância σ_0^2.

A estimação de modelos como o (7.5)[70] para "pares" de meses subsequentes incluídos no período que vai de novembro de 2006 a setembro de 2007, leva às estimativas do parâmetro β_1 para cada mês, cuja exponencial gera os índices de Jevons para os meses analisados. Como o índice de Jevons, diferentemente do índice de Dutot, não é afetado pela heterogeneidade encontrada na amostra, não se espera que os resultados de suas estimativas obtidos pela equação 7.3 sejam distintos dos oriundos dos ajustes da equação 7.5.

[70] Tentou-se também modelar *os erros* por uma função g, tal que $g(P_i, M_i, \delta) = \delta_{PMi}$, onde *P* e *M* são as variáveis de estratificação *Porte* e *Marca* e δ é um parâmetro de variância, de modo que Var (ε_i^τ) = $\sigma^2 \delta_{PM_i}^2$. Os resultados não foram satisfatórios.

Substituindo na equação 7.1 var(e_t) e var(e_{t-1}) pela variância dos preços livre da influência de marcas e estabelecimentos comerciais, ξ_t^2 e ξ_{t-1}^2, calculados por 7.5, de forma que:

$$I_J(p^{t-1}, p^t) \approx I_D(p^{t-1}, p^t)[1 + (1/2)\,\mathrm{var}(\xi^{t-1}) - (1/2)\,\mathrm{var}(\xi^t)] \qquad (7.6)$$

chega-se à aproximação entre o índice de Dutot, agora estimado livre da influência de locais e marcas, e o índice de Jevons. A Tabela (7.2), a seguir, apresenta as estimativas de exp(β_1), as variâncias sem os efeitos dos estabelecimentos e das marcas, os índices de Dutot calculados de forma simples e com heterogeneidade controlada a partir dos modelos hedônicos e a relação entre esses indicadores e o índice de Jevons. Os resultados mostram que, sem o efeito da variabilidade induzida por estabelecimentos comerciais e por variedade de marcas, o índice de Dutot se aproxima do índice de Jevons, como esperado teoricamente[71].

Tabela 7.2: Índices de Dutot e Jevons com heterogeneidade controlada, variâncias e estimativas da relação entre eles.

Mês	*Ijevons* exp(β_1)	ξ_{t-1}^2	ξ_t^2	*Idutot*	*Idutot.ch*	*Idutot/ Ijevons*	*Idutot.ch/ Ijevons*
nov/06	1,04011	*	0,00319	1,0405	1,04011	1,0003	*
dez/06	1,04534	0,00319	0,00335	1,0446	1,04542	0,9993	1,00008
jan/07	1,01149	0,00335	0,00342	1,0109	1,01151	0,9994	1,00003
fev/07	0,98900	0,00342	0,00233	0,9896	0,98846	1,0006	0,99946
mar/07	0,98681	0,00233	0,00344	0,9884	0,98736	1,0016	1,00055
abr/07	0,96927	0,00344	0,00415	0,9701	0,96961	1,0008	1,00035
mai/07	1,01388	0,00415	0,00409	1,0134	1,01384	0,9995	0,99997
jun/07	0,99647	0,00409	0,00306	0,9964	0,99595	0,9999	0,99948
jul/07	0,98499	0,00306	0,00326	0,9855	0,98508	1,0005	1,00010
ago/07	1,02753	0,00326	0,00238	1,0269	1,02708	0,9993	0,99956
set/07	1,05057	0,00238	0,00327	1,0509	1,05103	1,0003	1,00044

Fonte: Construção do autor.

*Como não havia mês anterior, não foi calculado o índice de outubro, o que também impossibilitou a obtenção dos resíduos em novembro (t-1).

[71] A não-resposta no banco original foi tratada com imputação pelo modelo.

A aproximação fica mais clara no resultado dos índices acumulados para 10 meses, de dezembro de 2006 a setembro de 2007, como mostra a Tabela 7.3.

Tabela 7.3 - Resumo dos resultados acumulados.

Índice	*Com controle da heterogeneidade*	*Simples*
Jevons	1,07446	1,07446
Dutot	1,07451	1,07596
Razão (Idutot/Ijevons)	1,000039	1,001394

Fonte: Construção do autor

Observa-se que a razão entre os índices de Jevons e de Dutot, quando este é controlado pela heterogeneidade, foi de 1,000039, enquanto que a razão dos índices sem controle foi de 1,001394.

A princípio, estas diferenças podem não parecer relevantes. Entretanto, vale destacar que o agregado elementar estudado, formado por especificações do subitem arroz de mesma gramatura (5kg), apresenta alto grau de homogeneidade. Mesmo assim, a regressão hedônica foi capaz de apontar e isolar as diferenças existentes entre marcas e estabelecimentos comerciais, propiciando estimativas das variações puras de preços.

8 – Comentários Finais

8.1 – Resumo

Neste trabalho, dados de pesquisas de índices de preços ao consumidor foram modelados num contexto longitudinal. Os resultados obtidos com a estimação dos modelos hierárquicos, capazes de captar a estrutura de covariância presente, apontaram para diferenças nos níveis e variabilidade dos preços em função de fatores como “porte” e “condição socioeconômica da região onde se localizam” os estabelecimentos comerciais, indicando que essas informações não devem ser descartadas nem no momento do dimensionamento amostral nem no processo de análise que antecede à divulgação dos índices mensais. Mesmo no capítulo cinco, onde as amostras eram aparentemente altamente homogêneas, contando com cotações de preços de apenas uma marca de arroz, as estimativas com base em verossimilhança captaram os efeitos de preditores nos preços médios e na dispersão.

Além de serem importantes instrumentos para conhecimento do fenômeno em estudo, os modelos hierárquicos podem ser adequados para o tratamento da não-resposta, muito comum em estudos longitudinais. No capítulo seis, onde foram apresentados os resultados alcançados com a simulação de diferentes métodos de imputação, mostrou-se que os resultados da modelagem longitudinal foram melhores que os demais. Foi possível perceber, ainda, como a imputação pela média e pelo preço do mês anterior conduziram a estatísticas viciadas. Já os métodos da modelagem longitudinal e o que contabiliza o preço observado anterior no cálculo da variação do bem, mas levando em conta a variação média do estrato, e que, ainda que num grau menos aprimorado, agrega informações transversais e longitudinais, apresentaram estimativas de vício praticamente nulas. No cálculo dos erros quadráticos médios, entretanto, diferiram, com o modelo longitudinal apresentando menores desvios.

Contudo, destaca-se, mais importante do que os resultados em si obtidos, deve ser a percepção tanto da característica longitudinal dos preços quanto do fato de que a dependência existente nos dados não deve ser desprezada na inferência.

A compreensão de como o conhecimento sobre a variância dos preços pode afetar a interpretação dos resultados foi complementada no sétimo capítulo com a avaliação da hipótese apresentada em Silver e Heravi (2007) de que parcela das diferenças entre os resultados fornecidos pelas aplicações das fórmulas de Jevons e de Dutot nos agregados elementares são influenciadas pela heterogeneidade das amostras. A estimação dos modelos hedônicos, que fornecem estimativas da dispersão dos preços livres dos efeitos dos estabelecimentos e das diferentes marcas, permitindo o cálculo da variância dos preços sem a interferência desses fatores, possibilitou o cálculo, para vários meses, do índice de Dutot, que não atende ao teste da comensurabilidade, com dispersão controlada. O resultado acumulado do índice de Dutot com esta característica foi, de fato, muito próximo do índice de Jevons.

Novamente, a homogeneidade da amostra estudada adquire importância nos comentários. No trabalho de Silver e Heravi (2007), a metodologia foi aplicada a uma base de dados composta por preços de produtos eletrônicos, cadastrados por códigos de barra (*scanner data*), e com grau de heterogeneidade muito acima da observada nos estratos aqui estudados. Ainda assim, os modelos aqui adotados foram capazes de depurar as estimativas, relevando a importância da técnica para captação de diferenças nos preços oriundas de alterações da qualidade.

8.2 – Limitações

Uma das limitações deste trabalho foi, portanto, o estudo baseado na amostra de apenas um subitem de consumo. Como o objetivo primeiro foi mostrar as diferentes frentes

em que poderiam ser utilizados os modelos longitudinais nas pesquisas de índices de preços, como predição e imputação, optou-se pelo foco nessas possibilidades de aplicações. Outra estratégia poderia ter sido, por exemplo, usar alguns métodos em um maior número de amostras ou estratos, comparando seus resultados.

Porém, pelo suporte teórico subjacente à modelagem longitudinal, onde, por exemplo, sustenta-se a robustez dos métodos de máxima verossimilhança para tratamento da não-resposta do tipo MAR, optou-se pelo privilégio da aplicação em pesquisas de IPC ao invés de um aprofundamento dos aspectos teóricos, já amplamente fundamentados na literatura.

8.3 – Trabalhos Futuros

Não obstante, seria interessante modelar uma amostra mais complexa de dados. O trabalho pode ser estendido não só para um grupo maior de subitens, que poderiam estar, inclusive, hierarquizados de acordo com e estrutura de classificação dos índices mas, também, para sub-regiões dentro de unidades federativas.

A avaliação mais pormenorizada da não-resposta encontrada em IPCs é outra vertente a ser seguida. No desenvolvimento dos modelos com efeitos aleatórios realizado nos capítulos cinco e sete, a suposição foi de não-resposta do tipo MAR. Se esta suposição não puder ser aceita, métodos mais robustos, como os modelos de seleção[72] (HECKMAN, 1976; AMEMIYA, 1984; HEDEKER e GIBBONS, 1997; DEMIRTAS, 2004) ou os modelos de mistura de padrão (LITTLE, 1995; HEDEKER e GIBBONS, 1997) podem ser técnicas mais apropriadas. Uma análise da aplicação destes métodos em IPC parece um caminho interessante a ser seguido.

[72] Em inglês, os modelos de seleção e de misturas de padrões são denominados, respectivamente, por *selection models* e *pattern mixture models*.

Os modelos de efeitos mistos apresentados no decorrer do texto trataram, sobretudo, da variância ou, em outras palavras, do controle de variáveis e características indutoras de variabilidade. Na prática, estes objetivos podem ser também alcançados recorrendo-se a planos amostrais complexos, pois no caso de desenhos estratificados, o controle da heterogeneidade pode decorrer diretamente do próprio desenho amostral, permitindo a obtenção de estimativas mais precisas[73]. Logo, outra importante questão a ser lembrada diz respeito à implementação de planos de amostragem nas pesquisas de índices de preços ao consumidor.

[73] Pessoa e Silva (1998) analisam dados amostrais complexos e sua importância.

9 – REFERÊNCIAS

AKAIKE, H. Information Theory and an Extension of the Maximum Likelihood Principle **Second International Symposium on Information Theory,** Em B. N. Petrov & F. Csaki (Editors) Akademiai Kiado, Budapest, 267-281, 1973.

AMEMIYA, T. Tobit Models: A Survey. **Journal of Econometrics**, 24, 3-61, 1984.

CARMO, H. C. E. **Índice de Preços ao Consumidor: Teoria e Análise de Modelos Factíveis Considerando as Bases de Dados Disponíveis**. Tese de livre Docência. Departamento de Economia da FEA-USP, 2004.

DEMIRTAS, H. Modeling Incomplete Longitudinal Data. **Journal of Modern Applied Statistical Methods**, Vol. 23, Nº. 2, 305-321, Novembro, 2004.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm**. Journal of the Royal Statistical Society**. Series B, Vol. 39, Nº. 1, 1-38, 1977.

DIEWERT, W. E**.** Axiomatic and Economic Approaches to Elementary Price Indexes. **National Bureau of Economic Researchs.** Cambridge. Working Paper Series. Nº. 5104, 27-29, 1995.

_____ **On the Stochastic Approach to Index Numbers.** University of British Columbia, Setembro, 1995.

_____ Methodological Problems with the Consumer Price Index1. Vancouver, Canada Department of Economics, University of British Columbia, V6T 1Z1. Dezembro, 2003.

DIEWERT, W. E.; NAKAMURA, A. O. **Essays in Index Number Theory**. Elsevier Science Publishers B.V, Vol. 1 1993.

DIGGLE, J. P.; LIANG, K. Y.; ZEGER, S. L. Analysis of Longitudinal Data**. Oxford Statistical Science Series** Nº. 13. Oxford Science Publications, 1994. 253 p.

EDGEWORTH, F. Y. The Doctrine of Index Numbers According to Mr. Correa Walsh. **The Economic Journal**, Vol. 11, 1923, p. 343–351.

ESTATÍSTICAS DO CADASTRO CENTRAL DE EMPRESAS 2006. Rio de Janeiro: IBGE, 2008, 167 p.

FERREIRA, S. G. **Inflação, Regras de Reajuste e Busca Sequencial: Uma Abordagem sob a Ótica da Dispersão de Preços Relativos**. Prêmio BNDES de Economia, Nº 19, Rio de Janeiro, 1995.

FISHER, I. **The Making of Index Numbers**. Boston, MA: Houghton-Mifflin, 1922.

HECKMAN, J. The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables, and a Simple Estimator for such Models. **Annals of Economic and Social Measurement**, 5, 475-492, 1976.

HEDEKER, D. R.; GIBBONS, R. D. Application of Random-effects Pattern-mixtre Models for Missing Data in Longitudinal Studies. **Psychological Methods**, 2, 64-78, 1997

_____ **Longitudinal Data Analysis**. Wiley Series in Probability and Statistics. Ed. Wiley. 2006, 337 p.

GUJARATI, D. N. **Econometria Básica**. São Paulo: Makron Books, 2000, 846 p.

ILO/IMF/OECD/UNECE/EUROSTAT/THE WORLD BANK. Consumer Price Index Manual: Theory and Practice. **International Labour Office**, Genebra, 2004.

KONÜS, A.A.;. The Problem of the True Index of the Cost of Living. **Econométrica,** 7, 10-29, 1939.

LAIRD, N. M. Missing Data in Longitudinal Studies. **Statistics in Medicine**, 7, 305-315, 1988.

LITTLE, R. J. A. Modeling the Drop-Out Mechanism in Repeated-Measures Studies. **Journal of the American Statistical Association**, 90, 1112-1121, 1995.

LITTLE, R. J. A.; RUBIN, D. **Statistical Analysis with Missing Data**. New York. Wiley, 1987.

MELO, F. A. M. Padrão de Vida, Custo de Vida e Índices de Preços ao Consumidor. **Revista Brasileira de Estatística**. Rio de Janeiro, Nº 37 (148), 445-456. Outubro/dezembro, 1976.

MOLENBERGHS, G. et al. Analyzing Incompolete Longitudinal Clinical Trial Data. **Biostatistics**, 5, 3, 445-464, 2004.

NATIONAL BUREAU OF ECONOMIC RESEARCH. **The Price Statistics of the Federal Government**, New York: Columbia University Press for the National Bureau of Economic Research, General Series, Nº 73, 1961.

PESSOA, D. G. C.; SILVA, P. L. N. **Análise de Dados Amostrais Complexos**. São Paulo: Associação Brasileira de Estatística, v.1, 1998, 187 p.

PINHEIRO, J. et al. Nlme**: Linear and Nonlinear Mixed Effects Models**. R Package Version 3.1 - 89, 2008.

PINHEIRO, J. C.; BATES, D. M. **Mixed-Effects Models in S and S-PLUS**. New York. Springer-Verlag, 2000, 526 p.

R DEVELOPMENT CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2008

RUBIN, D. B. Inference and Missing data. **Biometrika**, Vol. 63, 3, 581-592, Dezembro, 1976,

SÁINZ, P.; MANUELITO, S. Precios relativos em América Latina em Períodos de Baja Inflación y Câmbios Estructurales**. Revista de la Cepal**. Santiago de Chile, Nº. 89, 103-131, Agosto, 2006

SCHAFER, J. L. Analysis of Incomplete Multivariate Data. **Monographs on Statistics and Applied Probability** 72. London. CHAPMAN & HALL, 1997, 430 p.

SILVER, M.; HERAVI, S. Why Price Index Number Formulae Differ: Economic Theory and Evidence on Price Dispersion. **International Working Group on Price Indices,** Seventh Meeting. Paris, 2003.

_____ Why Elementary Price Index Number Formulas Differ: Evidence on Price Dispersion. **Journal of Econometrics**, 140, 874–883, 2007.

SINGER, J. D.; WILLETT, J. B. **Applied Longitudinal Data Analysis: modeling change and event occurrence**. New York. Oxford University Press. 2003, 644 p.

SISTEMA NACIONAL DE ÍNDICES DE PREÇOS AO CONSUMIDOR: **Estruturas de Ponderação a Partir da Pesquisa de Orçamentos Familiares 2002/2003 / IBGE**, Coordenação de Índices de Preços, Rio de Janeiro: IBGE, 2005.

_____ **Métodos de Cálculo**. Coordenação de Índices de Preços. Rio de Janeiro. 5ª Edição, Vol. 14, IBGE, 2007.

SNIJDERS, T. A. B.; BOSKER, R. J. **An Introduction to Basic and Advanced Multilevel Model**. London. SAGE Publications, 1999, 266 p.

STANFORD Encyclopedia of Philosophy. Disponível em http://plato.stanford.edu/ entries/ william-jevons/#Bio em 10/03/2009.

THE SWEDISH CONSUMER PRICE INDEX. **A handbook of methods**. Statistics Sweden, 2001. ISBN 91-618-1097-5.

UNITED Nations. Methods and Classifications. 2009. Disponível em http://unstats.un.org/ unsd/cr/registry/regct.asp?Lg=1 em 20/04/2009.

UNITED Nations Publication (St/ESA/STAT/SER.M/4/Rev.3), Sales No. E.90XVII.11. Disponível em http://www.ilo.org/public/english/bureau/stat/class/isic.htm em 05/06/2009.

VARIAN, H. **Microeconomic Analysis**, 3rd Edition. New York. W.W. Norton, 1993.

VELDE, F. R. **The Life and Times of Nicolas Dutot**. Chicago. Federal Reserve Bank of Chicago. Abril, 2009

Anexo 1 – Exemplo de Estrutura de Pesos Utilizada em IPC

IBGE
SISTEMA NACIONAL DE ÍNDICES DE PREÇOS AO CONSUMIDOR
Estrutura de Pesos para o IPCA
Referência: Janeiro de 2003
RIO DE JANEIRO

Código	Descrição	Peso (%)
0000000	INDICE GERAL	100,00
1000000	ALIMENTAÇÃO E BEBIDAS	22,59
1100000	ALIMENTAÇÃO NO DOMÍCILIO	15,07
1101000	CEREAIS, LEGUMINOSAS E OLEA	1,01
1101002	ARROZ	0,67
1101052	FEIJAO PRETO	0,34
1102000	FARINHAS,FÉCULAS E MASSAS	0,52
1102006	MACARRAO	0,26
1102013	FARINHA VITAMINADA	0,06
1102023	FARINHA DE MANDIOCA	0,07
1102029	MASSA SEMI-PREPARADA	0,14
1103000	TUBÉRCULOS,RAÍZES E LEGUMES	0,68
1103003	BATATA INGLESA	0,27
1103021	CHUCHU	0,06
1103028	TOMATE	0,14
1103043	CEBOLA	0,11
1103044	CENOURA	0,10
1104000	AÇÚCARES E DERIVADOS	0,80
1104003	ACUCAR REFINADO	0,48
1104023	CHOCOLATE EM BARRA	0,07
1104032	SORVETE	0,10
1104052	CHOCOLATE E ACHOCOLATADO EM PO	0,15
1105000	HORTALIÇAS EVERDURAS	0,15
1105001	ALFACE	0,07
1105005	COUVE	0,02
1105006	COUVE FLOR	0,02
1105010	REPOLHO	0,02
1105012	CHEIRO VERDE	0,02
1106000	FRUTAS	0,84
1106008	BANANA PRATA	0,20
1106017	MACA	0,16
1106018	MAMAO	0,08
1106019	MANGA	0,04
1106021	MELANCIA	0,04
1106023	PERA	0,06
1106027	TANGERINA	0,05
1106028	UVA	0,09
1106039	LARANJA PERA	0,14
1107000	CARNES FRESCAS E VÍSCERAS	1,65
1107018	CARNE DE PORCO	0,16
1107084	CONTRAFILE	0,30
1107085	FILE MIGNON	0,07
1107087	CHA DE DENTRO	0,09
1107088	ALCATRA	0,48
1107089	PATINHO	0,10
1107095	ACEM	0,26
1107099	COSTELA	0,18
1108000	PESCADOS	0,30
1108004	PEIXE-CORVINA	0,11
1108009	PEIXE-PESCADINHA	0,06
1108012	PEIXE-SARDINHA	0,02
1108013	CAMARAO	0,06
1108025	PEIXE LINGUADO	0,05
1109000	CARNES E PEIXES INDUSTRIALZADOS	0,77

IBGE
SISTEMA NACIONAL DE ÍNDICES DE PREÇOS AO CONSUMIDOR
Estrutura de Pesos para o IPCA
Referência: Janeiro de 2003
RIO DE JANEIRO - Continuação

Código	Descrição	Peso (%)
1109002	PRESUNTO	0,18
1109007	SALSICHA E SALSICHÃO	0,10
1109008	LINGUICA	0,23
1109023	BACALHAU	0,13
1109056	CARNE SECA	0,12
1110000	AVES E OVOS	1,15
1110009	FRANGO INTEIRO	0,39
1110010	FRANGO EM PEDAÇOS	0,58
1110044	OVO DE GALINHA	0,17
1111000	LEITES E DERIVADOS	2,18
1111004	LEITE DE PASTEURIZADO	0,86
1111008	LEITE CONDENSADO	0,08
1111009	LEITE EM PO	0,36
1111011	QUEIJO	0,61
1111019	IOGURTE	0,20
1111031	MANTEIGA	0,07
1112000	PANIFICADOS	2,13
1112003	BISCOITO DOCE	0,60
1112015	PAO FRANCES	1,28
1112018	PAO DE FORMA	0,25
1113000	ÓLEOS E GORDURAS	0,62
1113013	OLEO DE SOJA	0,31
1113014	AZEITE DE OLIVA	0,11
1113040	MARGARINA VEGETAL	0,21
1114000	BEBIDAS E INFUSÕES	1,68
1114001	SUCO DE FRUTAS	0,21
1114022	CAFE MOIDO	0,29
1114083	REFRIGERANTE E ÁGUA MINERAl	0,82
1114084	CERVEJA	0,36
1115000	ENLATADOS E CONSERVAS	0,18
1115006	ERVILHA EM CONSERVA	0,03
1115016	PALMITO EM CONSERVA	0,03
1115039	SARDINHA EM CONSERVA	0,03
1115057	AZEITONA	0,04
1115059	COGUMELO EM CONSERVA	0,02
1115075	ATUM EM CONSERVA	0,03
1116000	SAL E CONDIMENTOS	0,41
1116005	ATOMATADO	0,16
1116010	ALHO	0,12
1116033	MAIONESE	0,08
1116048	CALDO CONCENTRADO	0,05
1200000	ALIMENTAÇÃO FORA DO DOMICÍLIO	7,52
1201000	ALIMENTAÇÃO FORA DO DOMICÍLIO	7,52
1201001	REFEIÇÃO	3,96
1201003	LANCHE	1,44
1201005	CAFÉ DA MANHÃ	0,13
1201007	REFRIGERANTE E ÁGUA MINERAL	0,88
1201009	CAFEZINHO	0,07
1201048	CERVEJA	0,53
1201049	CHOPP	0,16
1201051	OUTRAS BEBIDAS ALCOOLICAS	0,13
1201061	DOCES	0,22
2000000	HABITAÇÃO	14,72

IBGE
SISTEMA NACIONAL DE ÍNDICES DE PREÇOS AO CONSUMIDOR
Estrutura de Pesos para o IPCA
Referência: Janeiro de 2003
RIO DE JANEIRO - Continuação

Código	Descrição	Peso (%)
2100000	ENCARGOS E MANUTENÇÃO	8,84
2101000	ALUGUEL E TAXAS	7,04
2101001	ALUGUEL RESIDENCIAL	3,07
2101002	CONDOMINIO	2,66
2101004	TAXA DE AGUA E ESGOTO	1,30
2103000	REPAROS	0,93
2103014	TINTA	0,16
2103032	AZULEJO E PISO	0,08
2103039	CIMENTO	0,13
2103042	MAO-DE-OBRA	0,55
2104000	ARTIGOS DE LIMPEZA	0,87
2104008	DETERGENTE	0,12
2104009	SABAO EM PO	0,40
2104012	DESINFETANTE	0,06
2104013	INSETICIDA	0,05
2104015	SABAO EM BARRA	0,10
2104020	LIMPADOR COM AMONIACO	0,06
2104032	AMACIANTE, ALVEJANTE	0,09
2200000	COMBUSTIVEIS E ENERGIA	5,88
2201000	COMBUSTÍVEIS (DOMÉSTICOS)	1,24
2201004	GAS DE BOTIJÃO	0,92
2201005	GAS ENCANADO	0,31
2202000	ENERGIA ELÉTRICA RESIDENCIAL	4,64
2202003	ENERGIA ELÉTRICA RESIDENCIAL	4,64
3000000	ARTIGOS DE RESIDÊNCIA	5,23
3100000	MÓVEIS E UTENSÍLIOS	2,24
3101000	MOBILIÁRIO	1,47
3101002	MÓVEL PARA SALA	0,60
3101003	MÓVEL PARA QUARTO	0,61
3101015	MÓVEL PARA COPA E COZINHA	0,18
3101017	COLCHAO	0,09
3102000	UTENSÍLIOS E ENFEITES	0,52
3102005	TAPETE	0,14
3102006	CORTINA	0,13
3102007	UTENSILIOS COPA E COZINHA DE METAL	0,07
3102009	UTENSÍLIOS COPA E COZ DE VIDRO E LOUÇA	0,05
3102010	UTENSILIOS DE PLASTICO	0,07
3102040	UTENSÍLIOS DIVERSOS	0,07
3103000	CAMA, MESA E BANHO	0,24
3103001	ROUPA DE CAMA	0,18
3103003	ROUPA DE BANHO	0,07
3200000	APARELHOS ELETROELETRONICOS	2,52
3201000	ELETRODOMÉSTICOS E EQUIPAMENTOS	0,97
3201001	REFRIGERADOR	0,39
3201002	CONDICIONADOR DE AR	0,13
3201006	MAQUINA DE LAVAR ROUPA	0,21
3201013	VENTILADOR	0,09
3201021	FOGAO	0,15
3202000	TV, SOM E INFORMATICA	1,55
3202001	TELEVISOR	0,54
3202003	APARELHO DE SOM	0,29
3202005	APARELHO DE DVD	0,20
3202028	MICROCOMPUTADOR	0,52
3300000	CONSERTOS E MANUTENÇÃO	0,47
3301000	CONSERTOS E MANUTENÇÃO	0,47
3301002	CONSERTO DE GELADEIRA E FREEZER	0,05

IBGE
SISTEMA NACIONAL DE ÍNDICES DE PREÇOS AO CONSUMIDOR
Estrutura de Pesos para o IPCA
Referência: Janeiro de 2003
RIO DE JANEIRO - Continuação

Código	Descrição	Peso (%)
3301006	CONSERTO DE TELEVISAO	0,12
3301009	CONSERTO DE APARELHO DE SOM	0,06
3301015	CONSERTO DE MAQUINA DE LAVAR/SECAR ROUPA	0,07
3301022	REFORMA DE ESTOFADO	0,17
4000000	VESTUÁRIO	4,82
4100000	ROUPAS	3,27
4101000	ROUPA MASCULINA	1,29
4101002	CALCA COMPRIDA MASCULINA	0,38
4101004	TERNO	0,09
4101006	SHORT E BERMUDA MASCULINA	0,21
4101009	CAMISA/CAMISETA MASCULINA	0,61
4102000	ROUPA FEMININA	1,28
4102002	CALCA COMPRIDA FEMININA	0,32
4102004	SAIA	0,14
4102005	VESTIDO	0,15
4102008	BLUSA	0,47
4102010	LINGERIE	0,12
4102013	BERMUDA E SHORT FEMININO	0,08
4103000	ROUPA INFANTIL	0,69
4103001	UNIFORME	0,05
4103002	CALCA COMPRIDA INFANTIL	0,09
4103007	VESTIDO INFANTIL	0,05
4103008	BERMUDA E SHORT INFANTIL	0,13
4103011	CAMISA/CAMISETA INFANTIL	0,22
4103017	FRALDA	0,14
4200000	CALÇADOS E ACESSÓRIOS	1,21
4201000	CALÇADOS E ACESSÓRIOS	1,21
4201002	SAPATO MASCULINO	0,20
4201003	SAPATO FEMININO	0,19
4201007	SANDALIA/CHINELO FEMININO	0,25
4201008	SANDALIA/CHINELO INFANTIL	0,08
4201015	BOLSA	0,14
4201063	TENIS	0,36
4300000	JÓIAS E BIJOUTERIAS	0,26
4301000	JÓIAS E BIJOUTERIAS	0,26
4301001	BIJUTERIA	0,11
4301002	JOIA	0,08
4301004	RELOGIO DE PULSO	0,08
4400000	TECIDOS E ARMARINHO	0,08
4401000	TECIDOS E ARMARINHO	0,08
4401001	TECIDO	0,06
4401002	ARTIGOS DE ARMARINHO	0,02
5000000	TRANSPORTE	20,53
5100000	TRANSPORTE	20,53
5101000	TRANSPORTE PÚBLICO	7,61
5101001	ONIBUS URBANO	4,15
5101002	TAXI	0,61
5101004	TREM	0,10
5101006	ONIBUS INTERMUNICIPAL	1,73
5101007	ONIBUS INTERESTADUAL	0,23
5101010	AVIAO	0,48
5101011	METRO	0,23
5101026	TRANSPORTE ESCOLAR	0,08
5102000	VEÍCULO PRÓPRIO	7,65

IBGE
SISTEMA NACIONAL DE ÍNDICES DE PREÇOS AO CONSUMIDOR
Estrutura de Pesos para o IPCA
Referência: Janeiro de 2003
RIO DE JANEIRO - Continuação

Código	Descrição	Peso (%)
5102001	AUTOMOVEL NOVO	3,16
5102004	EMPLACAMENTO E LICENÇA	0,76
5102005	SEGURO VOLUNTARIO DE VEICULO	0,44
5102009	ACESSÓRIOS E PEÇAS	0,60
5102010	PNEU E CAMARA DE AR	0,11
5102011	CONSERTO DE AUTOMOVEL	1,04
5102013	ESTACIONAMENTO	0,20
5102015	PEDAGIO	0,13
5102019	LAVAGEM E LUBRIFICACAO	0,16
5102020	AUTOMÓVEL USADO	0,80
5102053	MOTOCICLETA	0,24
5104000	COMBUSTÍVEIS(VEÍCULOS)	5,27
5104001	GASOLINA	4,55
5104002	ALCOOL	0,48
5104005	GAS VEICULAR	0,25
6000000	SAÚDE E CUIDADOS PESSOAIS	10,98
6100000	PRODUTOS FARMACÊUTICOS E ÓTICOS	3,63
6101000	PRODUTOS FARMACÊUTICOS	3,22
6101001	ANTIINFECCIOSO E ANTIBIOTICO	0,23
6101002	ANALGESICO E ANTITERMICO	0,20
6101003	ANTIINFLAMATORIO E ANTI-REUMATICO	0,36
6101004	ANTIGRIPAL E ANTITUSSIGENO	0,10
6101006	ANTIMICOTICO E PARASITICIDA	0,15
6101007	ANTIALERGICO E BRONCODILATADOR	0,23
6101009	GASTROPROTETOR	0,18
6101010	VITAMINA E FORTIFICANTE	0,21
6101011	HORMONIO	0,31
6101013	PSICOTRÓPICO E ANOREXÍGENO	0,42
6101014	HIPOTENSOR E ANTICOLESTERINICO	0,71
6101051	OFTALMOLOGICO	0,12
6102000	ÓCULOS E LENTES	0,41
6102001	LENTE DE GRAU	0,30
6102002	ARMACAO DE OCULOS	0,11
6200000	SERVIÇOS DE SAUDE	5,00
6201000	SERVIÇOS MÉDICOS E DENTÁRIOS	1,40
6201002	MEDICO	0,39
6201003	DENTISTA	0,71
6201005	APARELHO DENTÁRIO	0,30
6202000	SERVIÇOS LABORATORIAIS E HOSPITALARES	0,76
6202004	HOSPITALIZAÇÃO E CIRURGIA	0,36
6202006	ELETRODIAGNOSTICO	0,11
6202008	ASILO	0,29
6203000	PLANO DE SAÚDE	2,85
6203001	PLANO DE SAUDE	2,85
6300000	CUIDADOS PESSOAIS	2,35
6301000	HIGIENE PESSOAL	2,35
6301001	PRODUTO PARA CABELO	0,27
6301006	PRODUTO PARA PELE	0,29
6301007	PRODUTO PARA HIGIENE BUCAL	0,16
6301010	PRODUTO PARA UNHA	0,10
6301011	PERFUME	0,89
6301014	DESODORANTE	0,10
6301015	ABSORVENTE HIGIENICO	0,09
6301016	SABONETE	0,15
6301017	PAPEL HIGIENICO	0,22
6301020	ARTIGO DE MAQUIAGEM	0,08

IBGE
SISTEMA NACIONAL DE ÍNDICES DE PREÇOS AO CONSUMIDOR
Estrutura de Pesos para o IPCA
Referência: Janeiro de 2003
RIO DE JANEIRO - Continuação

Código	Descrição	Peso (%)
7000000	DESPESAS PESSOAIS	8,08
7100000	SERVIÇOS PESSOAIS	4,09
7101000	SERVIÇOS PESSOAIS	4,09
7101005	MANICURE E PEDICURE	0,41
7101008	BARBEIRO	0,12
7101009	CABELEIREIRO	0,86
7101010	EMPREGADO DOMÉSTICO	2,12
7101076	SERVIÇO BANCARIO	0,58
7200000	RECREAÇÃO, FUMO E FILMES	3,99
7201000	RECREAÇÃO	2,85
7201001	CINEMA	0,33
7201006	CLUBE	0,16
7201008	DISCO LASER	0,22
7201010	INSTRUMENTO MUSICAL	0,11
7201019	BICICLETA	0,09
7201020	ALIMENTOSPARA ANIMAIS	0,28
7201023	BRINQUEDOS	0,43
7201052	ALUGUEL DE DVD E FITA VIDEOCASSETE	0,16
7201054	BOITE,DANCETERIA E DISCOTECA	0,22
7201063	JOGOS DE AZAR	0,69
7201090	HOTEL	0,17
7202000	FUMO	0,90
7202041	CIGARRO	0,90
7203000	FOTOGRAFIA E FILMAGEM	0,24
7203003	REVELACAO E COPIA	0,24
8000000	EDUCAÇÃO	6,48
8100000	CURSOS, LEITURA E PAPELARIA	6,48
8101000	CURSOS	4,07
8101001	CRECHE	0,31
8101002	EDUCAÇÃO INFANTIL	0,34
8101003	ENSINO FUNDAMENTAL	1,11
8101004	ENSINO MÉDIO	0,56
8101005	ENSINO SUPERIOR	1,48
8101006	POS-GRADUACAO	0,27
8102000	LEITURA	1,15
8102001	JORNAL DIÁRIO	0,42
8102002	ASSINATURA DE JORNAL	0,11
8102004	REVISTA	0,35
8102005	LIVRO	0,27
8103000	PAPELARIA	0,35
8103001	CADERNO	0,09
8103002	FOTOCOPIA	0,08
8103014	ARTIGOS DE PAPELARIA	0,17
8104000	CURSOS DIVERSOS	0,92
8104001	CURSO PREPARATORIO	0,12
8104003	CURSO DE IDIOMA	0,46
8104004	CURSO DE INFORMATICA	0,14
8104006	GINASTICA	0,19
9000000	COMUNICAÇÃO	6,55
9100000	COMUNICAÇÃO	6,55
9101000	COMUNICAÇÃO	6,55
9101001	CORREIO	0,12
9101002	TELEFONE FIXO	3,82
9101003	TELEFONE PUBLICO	0,26
9101008	TELEFONE CELULAR	1,28
9101010	TV A CABO	0,66

IBGE
SISTEMA NACIONAL DE ÍNDICES DE PREÇOS AO CONSUMIDOR
Estrutura de Pesos para o IPCA
Referência: Janeiro de 2003
RIO DE JANEIRO - Continuação

Código	**Descrição**	**Peso (%)**
9101018	ACESSO A INTERNET	0,10
9101019	APARELHO TELEFONICO	0,32

Anexo 2 – Resultados das Regressões Logísticas

Resultados de regressões logísticas para avaliar indícios de diferenças nas probabilidades de não-resposta em função do Porte do estabelecimento comercial:

	Coeficiente	***Estimativa***	***p-valor***
PRODUTO 02000	Intercepto	0,1143	0,017
	Porte (MP)	0,1831	6×10^{-9}
PRODUTO 5000	Intercepto	-2,2824	$< 2 \times 10^{-16}$
	Porte (MP)	0,5657	0,0072
PRODUTO 07000	Intercepto	-2,6810	$2,3 \times 10^{-16}$
	Porte (MP)	1,5824	$2,5 \times 10^{-5}$
PRODUTO 09000	Intercepto	1,3568	$< 2 \times 10^{-16}$
	Porte (MP)	0,5489	0,0179
PRODUTO 10000	Intercepto	-1,6514	$< 2 \times 10^{-16}$
	Porte (MP)	1,1049	$3,82 \times 10^{-6}$

Anexo 3 – "Scripts" do R para Validação Cruzada

```
#Criação dos objetos para recebimento das estatísticas:

#Erros quadráticos médios para cada amostra

  EQM.modelo<-numeric(10)
  EQM.media.mar<-numeric(10)
  EQM.mant.Rel<-numeric(10)
  EQM.mant.MAR<-numeric(10)

#Média dos preço imputados:

  p.medio.imp.modelo<-numeric(10)
  p.medio.imp.media<-numeric(10)
  p.medio.imp.varmedia<-numeric(10)
  p.medio.imp.ult<-numeric(10)

#Preço médios observados

  p.medio.obs.modelo<-numeric(10)
  p.medio.obs.media<-numeric(10)
  p.medio.obs.varmedia<-numeric(10)
  p.medio.obs.ult<-numeric(10)

 #Vício absoluto (Diferença entre preços médios imputados e observados)

  dif.p.medios.modelo<-numeric(10)
  dif.p.medios.media<-numeric(10)
  dif.p.medios.varmedia<-numeric(10)
  dif.p.medios.ult<-numeric(10)

 #Vício relativo global - [soma(|yi-yobs|/yobs)]/N

  vicio.global.modelo<-numeric(10)
  vicio.global.media<-numeric(10)
  viciorelativo.global.mant.rel<-numeric(10)
  viciorelativo.global.ult.obs<-numeric(10)

 #Vício relativo - soma(|yi-yobs|/yobs)

  viciorelativo.modelo<-numeric(10)
  vicio.media.abs<-numeric(10)
  viciorelativo.mant.Rel<-numeric(10)
  viciorelativo.ult.obs<-numeric(10)
```

```
#Seleção das amostras considerando-se que a não-resposta é mar e  construção da #função
#"Tabela.Imput" para gerar os resultados a partir das suposições a respeito das
#probabilidades de não-resposta para estabelecimentos  de grande porte
#e estabelecimentos de porte pequeno ou médio.

 Tabela.Imput<-function(P.nr.G,P.nr.MP){

#Criação d a variável que representa a probabilidade de não-resposta por tipo
#de local no data frame que contém os dados (dados1_prod05s)

dados1_prod05s$prob_nresposta<-numeric(785)

#Atribuição das probabilidades

dados1_prod05s$prob_nresposta[dados1_prod05s$Porte2=="G"]<-P.nr.G
dados1_prod05s$prob_nresposta[dados1_prod05s$Porte2=="MP"]<-P.nr.MP

#Inclusão da var auxiliar de ID ("ident") no data frame

 dados1_prod05s$ident<-(1:785)

#Criação das variáveis que receberão os id de cada amostra
#Como serão 10 amostras, são criadas as variáveis "amostra1" a "amostra10"

dados1_prod05s$amostra1<-numeric(785)
dados1_prod05s$amostra2<-numeric(785)
dados1_prod05s$amostra3<-numeric(785)
dados1_prod05s$amostra4<-numeric(785)
dados1_prod05s$amostra5<-numeric(785)
dados1_prod05s$amostra6<-numeric(785)
dados1_prod05s$amostra7<-numeric(785)
dados1_prod05s$amostra8<-numeric(785)
dados1_prod05s$amostra9<-numeric(785)
dados1_prod05s$amostra10<-numeric(785)

#Procedimento para seleção de cada amostra usando a sampling:

library(sampling)

#Para amostra número 1

name=dados1_prod05s$ident
set.seed(1000+1)      para amostra1
n=80
pikpoisson=inclusionprobabilities(dados1_prod05s$prob_nresposta,n)
```

```
#Observar que a probabilidade de inclusão na amostra para validação cruzada
#de estabelecimentos de grande porte é menor do que a de estabelecimentos
#de porte pequeno ou médio, Isto se deve ao fato de que esta amostra é
#para estimativa dos EQMs. A suposição é de perdas maiores em estabelecimentos
#de menor porte.


#Seleção da amostra

s=UPpoisson(pikpoisson)

#Amostra selecionada

id_amostra<-getdata(name,s)
id_amostra

#Criação da variável comum para merge dos data.frame

id_amostra$ident<-id_amostra[,"ID_unit"]

#Juntando os dois data frame

dados1_prod05s_Poisson<-
merge(dados1_prod05s,id_amostra,by="ident",all.x=TRUE)

#Levando os dados para a var "amostra1"

dados1_prod05s_Poisson[,36]<-dados1_prod05s_Poisson[,47]

#Alteração os NAs para zeros

dados1_prod05s_Poisson[,36][is.na(dados1_prod05s_Poisson[,47])]<-0

#Sub-rotina para as amostras 2 a 10

for (r in 37:45)
   {
      set.seed(1000+r)      para amostraj j=r-32
      n=80
      pikpoisson=inclusionprobabilities(dados1_prod05s$prob_nresposta,n)

#Seleção

      s=UPpoisson(pikpoisson)

#A amostra é:

      id_amostra<-getdata(name,s)
```

```
#Criação da  variável comum para merge (usar a variável "ident")

        id_amostra$ident<-id_amostra[,"ID_unit"]

#Merge

       dados1_prod05s_Poisson<-merge(dados1_prod05s_Poisson,id_amostra,
        by="ident",all.x=TRUE)

#Levando os dados para a var "amostra j"

       dados1_prod05s_Poisson[,r]<-dados1_prod05s_Poisson[,48]

 #Transformação dos NAs em zeros
       dados1_prod05s_Poisson[,r][is.na(dados1_prod05s_Poisson[,r])]<-0

#Exclusão das colunas ID_unit.y(48) e data.y(49)

      dados1_prod05s_Poisson<-dados1_prod05s_Poisson[,-(48:49)]
  }

#Procedimento para "Varmedia"
#Procedimento para incorporar no dados1.mant.Rel às amostras para cálculo dos
#EQMs.

 dados1_prod05s_Poisson.ord<-dados1_prod05s_Poisson
 [order(dados1_prod05s_Poisson$time,dados1_prod05s_Poisson$LOCAL),]
 Mergex<-merge(dados1.mant.Rel.ord,dados1_prod05s_Poisson.ord, all.x=T)
 unique(Mergex$amostra1)
 Mergex[Mergex$amostra1==19,"VAL"]
 dados1_prod05s_Poisson[dados1_prod05s_Poisson$amostra1==19,"VAL"]

#Cálculo dos Relativos por produto em cada mês: obs.: mes 1 = 1.

Mergex$Rel.Produto<-numeric(888)

Mergex$Rel.Produto[Mergex$time==1]<-1

  for (j in 2:12)

    {

      for (i in 1:888)
          {
            Mergex$Rel.Produto[Mergex$time==j]<-
            (Mergex$Media_mes[Mergex$time==j]/Mergex$Media_mes
            [Mergex$time==j-1])
          }
     }
```

```
#Cálculo dos imputados: Rel.Produto x imputado.mant.

Mergex$imputado.mant<-Mergex$Rel.Produto*Mergex$Media_mes

#Procedimento para último preço observado.

dados1_prod05s_Poisson.ord<-
dados1_prod05s_Poisson[order(dados1_prod05s_Poisson$time,
dados1_prod05s_Poisson$LOCAL),]
imputado.mant<-dados1.mes.anterior$imputado.mant
dados1_prod05s_Poisson.ord$imputado.mant<-imputado.mant

for (z in 36:45)
  {

 #Média

  mod05.media.mar=lm(VAL~factor(time),
  data=dados1_prod05s_Poisson[dados1_prod05s_Poisson[,z]==0,])
  pred_mod05.media.mar=predict(mod05.media.mar,
  newdata=dados1_prod05s_Poisson[dados1_prod05s_Poisson[,z]!=0,])
  observed.mar=dados1_prod05s_Poisson[dados1_prod05s_Poisson[,z]!=0,"VAL"]
  EQM.media.mar[z-35]=mean((pred_mod05.media.mar-observed.mar)^2)
  vicio.media.abs[z-35]=sum(pred_mod05.media.mar-observed.mar)/
  length(observed.mar)
  vicio.media.mar[z-35]=sum(pred_mod05.media.mar-observed.mar)
        viciorelativo.media[z-35]<-sum(
        (abs(pred_mod05.media.mar-observed.mar))/observed.mar)
        vicio.global.media[z-35]<-viciorelativo.media[z-35]/
        length(observed.mar)
        p.medio.imp.media[z-35]<-mean(pred_mod05.media.mar)
        p.medio.obs.media[z-35]<-mean(observed.mar)
        dif.p.medios.media[z-35]<- p.medio.imp.media
        [z-35]-p.medio.obs.media[z-35]

 #Modelo

 library(nlme)
  mod0.ex.z=update(mod05.pol2.varIdent.PN, cor=corARMA(p=3,q=0),
  data=dados1_prod05s_Poisson[dados1_prod05s_Poisson[,z]==0,])
  pred.mod.mar=predict(mod0.ex.z,
  newdata=dados1_prod05s_Poisson[dados1_prod05s_Poisson[,z]!=0,])
  observed.mar=dados1_prod05s_Poisson[dados1_prod05s_Poisson[,z]!=0,"VAL"]
  EQM.modelo[z-35]=mean((pred.mod.mar-observed.mar)^2)
        viciorelativo.modelo[z-35]<-sum(
        (abs(pred.mod.mar-observed.mar))/observed.mar)
        vicio.global.modelo[z-35]<-viciorelativo.modelo[z-35]/
        length(observed.mar)
```

```
p.medio.imp.modelo[z-35]<-mean(pred.mod.mar)
p.medio.obs.modelo[z-35]<-mean(observed.mar)
dif.p.medios.modelo[z-35]<- p.medio.imp.modelo
[z-35]-p.medio.obs.modelo[z-35]
```

#Varmedia

```
EQM.mant.Rel[z-35]<-mean((Mergex$imputado.mant
[Mergex[,(z+3)]!=0&!is.na(Mergex[,(z+3)])]-
(Mergex$VAL[Mergex[,(z+3)]!=0&!is.na(Mergex[,(z+3)])]))^2)
viciorelativo.mant.Rel[z-35]<-sum((abs(Mergex$imputado.mant
[Mergex[,(z+3)]!=0&!is.na(Mergex[,(z+3)])]-
Mergex$VAL[Mergex[,(z+3)]!=0&!is.na(Mergex[,(z+3)])]))/
Mergex$VAL[Mergex[,(z+3)]!=0&!is.na(Mergex[,(z+3)])])
L<-length(Mergex$VAL[Mergex[,(z+3)]!=0&!is.na(Mergex[,(z+3)])])
viciorelativo.global.mant.rel[z-35]<-viciorelativo.mant.Rel[z-35]/L
p.medio.imp.varmedia[z-35]<-mean(Mergex$imputado.mant
[Mergex[,(z+3)]!=0&!is.na(Mergex[,(z+3)])])
p.medio.obs.varmedia[z-35]<-mean
(Mergex$VAL[Mergex[,(z+3)]!=0&!is.na(Mergex[,(z+3)])])
dif.p.medios.varmedia[z-35]<- p.medio.imp.varmedia
[z-35]-p.medio.obs.varmedia[z-35]
```

#Último preço observado

```
EQM.mant.MAR[z-35]<-mean(((dados1_prod05s_Poisson.ord[,"VAL"]-
dados1_prod05s_Poisson.ord[,"imputado.mant"])
[dados1_prod05s_Poisson.ord[,z]>0])^2)
viciorelativo.ult.obs[z-35]<-sum(
(abs((dados1_prod05s_Poisson.ord[,"imputado.mant"]-
dados1_prod05s_Poisson.ord[,"VAL"])[dados1_prod05s_Poisson.ord[,z]>0]))/
(dados1_prod05s_Poisson.ord[,"VAL"])[dados1_prod05s_Poisson.ord[,z]>0])
L=length((dados1_prod05s_Poisson.ord[,"VAL"])
[dados1_prod05s_Poisson.ord[,z]>0])
viciorelativo.global.ult.obs[z-35]<-viciorelativo.ult.obs[z-35]/L
p.medio.imp.ult[z-35]<-mean((dados1_prod05s_Poisson.ord[,"imputado.mant"])
[dados1_prod05s_Poisson.ord[,z]>0])
p.medio.obs.ult[z-35]<-mean((dados1_prod05s_Poisson.ord[,"VAL"])
[dados1_prod05s_Poisson.ord[,z]>0])
dif.p.medios.ult[z-35]<-p.medio.imp.ult[z-35]-p.medio.obs.ult[z-35]
```

#Resultados

```
Tabela.Result.media<-data.frame
("EQM.media"=EQM.media.mar,"p.medio.imp.media"=p.medio.imp.media,
"p.medio.obs.media"=p.medio.obs.media,"Dif.media"=dif.p.medios.media,
"vicio.global.media"=vicio.global.media)
Tabela.Result.geral.media<-c("EQM.media.g"=mean(EQM.media),
```

```
"p.medio.imp.media.g"=mean(p.medio.imp.media),
"p.medio.obs.media.g"=mean(p.medio.obs.media),
"Dif.media.g"=mean(dif.p.medios.media),
"vicio.global.media.g"=mean(vicio.global.media))

Tabela.Result.modelo<-data.frame("EQM.modelo"=EQM.modelo,
"p.medio.imp.modelo"=p.medio.imp.modelo,
"p.medio.obs.modelo"=p.medio.obs.modelo,"Dif.modelo"=dif.p.medios.modelo,
"vicio.global.modelo"=vicio.global.modelo)
Tabela.Result.geral.modelo<-c("EQM.modelo.g"=mean(EQM.modelo),
"p.medio.imp.modelo.g"=mean(p.medio.imp.modelo),
"p.medio.obs.modelo.g"=mean(p.medio.obs.modelo),"Dif.modelo.g"=
mean(dif.p.medios.modelo),
"vicio.global.modelo.g"=mean(vicio.global.modelo))

Tabela.Result.varmedia<-data.frame("EQM.Varmedia"=EQM.mant.Rel,
"p.medio.imp.varmedia"=p.medio.imp.varmedia,
"p.medio.obs.varmedia"=p.medio.obs.varmedia,
"Dif.varmedia"=dif.p.medios.varmedia,
"vicio.global.varmedia"=viciorelativo.global.mant.rel)
Tabela.Result.geral.varmedia<-c("EQM.Varmediamodelo.g"=mean(EQM.mant.Rel),
"p.medio.imp.varmedia.g"=mean(p.medio.imp.varmedia),
"p.medio.obs.varmedia.g"=mean(p.medio.obs.varmedia),
"Dif.varmedia.g"=mean(dif.p.medios.varmedia),
"vicio.global.varmedia.g"=mean(viciorelativo.global.mant.rel))

Tabela.Result.ult.obs<-data.frame("EQM.ult.obs"=EQM.mant.MAR,
"p.medio.imp.ult.obs"=p.medio.imp.ult,
"p.medio.obs.ult.obs"=p.medio.obs.ult,"Dif.ult.obs"=dif.p.medios.ult,
"vicio.global.ult.obs"=viciorelativo.global.ult.obs)
Tabela.Result.geral.ult.obs<-c("EQM.ult.obs.g"=mean(EQM.mant.MAR),
"p.medio.imp.ult.obs.g"=mean(p.medio.imp.ult),
"p.medio.obs.ult.obs.g"=mean(p.medio.obs.ult),
"Dif.ult.obs.g"=mean(dif.p.medios.ult),
"vicio.global.ult.obs.g"=mean(viciorelativo.global.ult.obs))

Tabela.Result.list<-list("Tab.Média"=Tabela.Result.media,
"Tab.Resumo.Geral.Média"=Tabela.Result.geral.media,
"Tab.Modelo"=Tabela.Result.modelo,
"Tab.Result.Geral.Modelo"=Tabela.Result.geral.modelo,
"Tab.Varmédia"=Tabela.Result.varmedia,
"Tab.Result.Geral.Varmedia"=Tabela.Result.geral.varmedia,
"Tab.Últ.obs"=Tabela.Result.ult.obs,
"Tabela.Result.Geral.Ult.Obs."=Tabela.Result.geral.ult.obs)

#Resultado geral tabelado

Tabela.Result.list
  }
```

```
#Fechamento da função
  }

#Observação de  resultados para algumas probabilidades

Tabela.Result.list.10.15<-Tabela.Imput(0.10,0.15)
Tabela.Result.list.30.30<-Tabela.Imput(0.30,0.30)
Tabela.Result.list1010<-Tabela.Imput(10,10)

#Determinação do vício absoluto para diversos valores
# da razão entre as probabilidades de não-resposta de acordo
# com o tipo de estabelecimento.

#Alocando espaço
viciomedia<-numeric(11)
viciomodelo<-numeric(11)
viciovarmedia<-numeric(11)
vicio.ult.obs<-numeric(11)

#Cálculo das estimativas

for (j in 10:20)
 {
  Tabela.result.list<-Tabela.Imput(10,j)
  viciomedia[j-9]<-Tabela.result.list[[2]][4]
  viciomodelo[j-9]<-Tabela.result.list[[4]][4]
  viciovarmedia[j-9]<-Tabela.result.list[[6]][4]
  vicio.ult.obs[j-9]<-Tabela.result.list[[8]][4]
 }

#Resultados
viciomedia
viciomodelo
viciovarmedia
vicio.ult.obs
```

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo