

**Centro de Estudos de Política e Economia do Setor Público (Cepesp – FGV)
Laboratório de Urbanismo da Metrópole (Lume – FAUUSP)**

Compatibilização entre tonelagem, valor monetário e a matriz de carga

Relatório de Pesquisa

São Paulo
Novembro
2006

Coordenador: Ciro Biderman

Pesquisadores: Henrique Eduardo F. Vinhais
Marcelo Tyszler

Introdução

Um dos grandes desafios desse projeto de pesquisa foi criar um sistema que compatibilizasse dados monetários com dados físicos. Ainda que a compatibilização seja em princípio simples (uma vez conhecidos os preços dos produtos) na prática a compatibilização é bastante complexa. Nessa pesquisa testamos diversos modelos alternativos que serão descritos sumariamente nesse relatório. Daremos maior atenção ao modelo adotado efetivamente. Acreditamos que as estimativas representem o melhor ajuste possível em função das informações atualmente disponíveis para consulta.

A principal fonte de informação primária para esse estudo, como usual nas análises regionais ou urbanas, foram os dados de emprego e massa salarial por município. Esses dados estão disponíveis anualmente e são compilados pelo ministério do trabalho e emprego (MTE). O problema, também bastante destacado na literatura, é que os dados do ministério referem-se apenas ao emprego formal. Em alguns países esse poderia ser considerado um problema menor. No entanto, não é esse o caso do Brasil onde o emprego formal representa uma parcela muito elevada do emprego total. Assim, na primeira parte desse relatório apresentamos a técnica utilizada para converter os dados formais em totais.

Em seguida apresentamos a estratégia adotada para obter preços médios por produto. Utilizamos todas as fontes disponíveis para esse fim e adotamos alguns critérios para definir qual seria a fonte mais confiável para esse fim. Na terceira seção apresentamos o modelo de aproximação de valores físicos (pesos) a partir de grandezas monetárias. Essencialmente estimamos a base de exportação a partir da produção e importação de uma região e, em seguida, a importação a partir do nível estimado para a exportação. Esses valores foram então ajustados por Box-Cox¹ pois apresentaram sinais claros de não linearidades.

Modelo de pesos

A informalidade no mercado de trabalho é um dos assuntos mais discutidos na sociedade atual. Embora a questão (ou mesmo a definição) da informalidade seja bastante complexa, ela é entendida, quase sempre, como o percentual de empregados sem carteira de trabalho assinada somado ao de trabalhadores por conta própria, em relação ao total de ocupados. No Brasil, segundo dados do IPEA (Instituto de Pesquisa Econômica Aplicada) o grau de informalidade está na ordem de 50%. A questão não está somente no número de trabalhadores com carteira assinada, amparados pela lei, e os que trabalham na informalidade. Na verdade, a discussão envolve o peso da informalidade na perda de competitividade internacional da economia brasileira.

A identificação das causas desta informalidade passa pela discussão de fatores como elevada carga tributária, rigidez da legislação trabalhista e características institucionais do Estado brasileiro, entre outros. Na busca por explicar o grau de informalidade da economia, este trabalho propõe uma estimativa do emprego e massa salarial total a partir de características regionais e da economia formal. Para tanto, parte do conceito de que o grau de informalidade pode ser explicado a partir de variáveis observáveis dos municípios e estados e informações do mercado de trabalho formal. Desta forma, é

¹ Essa sugestão foi realizada por Bernardo Alvim ao qual agradecemos valendo as isenções de praxe.

possível calcular uma ponderação que, aplicada sobre o emprego e massa salarial formal, nos permite encontrar o emprego e massa salarial total da economia.

Base de Dados

Para este trabalho, são necessárias informações sobre o mercado de trabalho formal e total (formal + informal), além de características locais, para os anos de 1991, 1996 e 2000. Para o mercado de trabalho formal, são utilizados os micro-dados da base de dados da RAIS (Relação Anual de Informações Sociais) do Ministério do Trabalho e Emprego. As informações das empresas são agregadas no nível município e setor. Assim, trabalhamos com as seguintes variáveis da RAIS:

- município em que a empresa se localiza;
- setor ao qual a empresa pertence;
- total de emprego formal das empresas do setor e município;
- total de massa salarial formal das empresas do setor e município;
- salário médio: total da massa salarial formal dividido pelo total do emprego formal;
- número de empresas: quantidade de empresas com emprego formal maior do que zero.

Já para o mercado de trabalho total e características locais, são utilizados os micro-dados do IBGE (Instituto Brasileiro de Geografia e Estatística) dos Censos Demográficos de 1991 e 2000 e da PNAD (Pesquisa Nacional por Amostra de Domicílios) de 1996. Assim como os dados da RAIS, o nível de agregação utilizado é município e setor, com as seguintes variáveis:

- município onde o indivíduo reside (Censo 1991 e PNAD 1996);
- município onde o indivíduo trabalha (apenas existente no Censo 2000);
- setor ao qual a empresa, em que o indivíduo trabalha, pertence;
- população urbana: total da população que reside em domicílios localizados em área urbana;
- total de estudantes: total da população com idade acima de 5 anos, freqüentando algum tipo de curso;
- população em idade ativa: total da população com idade acima de 10 anos;
- população masculina: total da população de sexo masculino;
- escolaridade média: quantidade média de anos de estudo dos chefes de domicílio;
- emprego total: soma do total de pessoas ocupadas (com informação de setor de atividade);
- massa salarial total: soma do total de rendimentos no trabalho principal com o total de rendimentos nos demais trabalhos.

Vale ressaltar que enfrentamos dois problemas com as informações disponíveis. Primeiramente, tanto para 1991 e 1996, não possuímos a informação de onde cada indivíduo trabalha. Por esta razão, este trabalho estima o emprego total a partir do emprego formal, considerando o perfil dos trabalhadores de 2000, em que temos a informação do local de trabalho do indivíduo. Além disso, em 1996, por trabalharmos com a PNAD, a maior desagregação possível é unidade da federação e região metropolitana, totalizando 36 *ufrmt*.

Classificações Setoriais

Um dos principais desafios enfrentados neste trabalho para a adoção da metodologia proposta é a questão das diversas classificações setoriais existentes, uma vez que cada uma das bases de dados utilizadas possui uma classificação própria.

Como ponto de partida, recorreremos à classificação de atividades econômicas da CONCLA (Comissão Nacional de Classificação) do IBGE, que apresenta a correspondência entre as classificações CNAE (Classificação Nacional de Atividades Econômicas) domiciliar, CNAE e setores do Censo de 1991. Como o Censo de 2000 utiliza a classificação CNAE Domiciliar, as RAIS de 1996 e 2000 utilizam a CNAE e a PNAD de 1996 utiliza a classificação do Censo de 1991, temos a correspondência entre os dois Censos, a PNAD e as duas RAIS citadas. Contudo, não temos nenhuma correspondência com a classificação setorial da RAIS de 1991.

Para a compatibilização entre CNAE's poderíamos recorrer às empresas existentes em 1991 na RAIS e que não fecharam entre 1991 e 1996. Assim, saberíamos a correspondência entre as classificações de 1991 e 1996. Entretanto, este recurso não apresenta bons resultados, pois as empresas de um mesmo setor em 1991 não necessariamente são classificadas num mesmo setor em 1996. Desta forma, ocorrem inúmeros casos em que empresas de um mesmo setor em 1991 vão para setores diferentes em 1996. Dado este problema, por simplificação, optamos por classificar os setores de acordo com o código e descrição de ambos os anos, tendo estas empresas apenas como auxílio.

Após a compatibilização de todas as classificações de setores, foi necessário agrupar os setores em duas classificações. Uma primeira, com a agregação das classificações CNAE Domiciliar, CNAE e da RAIS de 1991, chamada de SCNAE (com 112 setores) e uma segunda classificação que concilia a SCNAE com a do Censo 1991, ou seja, que concilia todas as classificações citadas, sendo chamada de SSCNAE (com 75 setores).

Modelo Econométrico

a) Modelo de Emprego

O modelo proposto utiliza as características dos municípios e do mercado de trabalho do ano de 2000 para estimar o emprego total, a partir da seguinte equação:

$$\ln e_{ij}^t = \beta_{0i} + \beta_{1ij} \ln X_j + \beta_{2ij} \ln e_{ij}^f + \varepsilon_{ij}, \quad (1)$$

onde e_{ij}^t é o emprego total do setor i no município j , β_{0i} é o termo constante da regressão de cada setor i , β_{1ij} e β_{2ij} são os coeficientes das variáveis explicativas da regressão de cada setor i , X_j é o conjunto de variáveis de características do município, e_{ij}^f é o conjunto de características do mercado de trabalho formal do setor i no município j e ε_{ij} é o resíduo da regressão.

Como pode ser observado, o modelo proposto trabalha com as variáveis em logaritmo natural.

É importante detalhar que as variáveis constantes em X_j são:

- população urbana (*popurb*);
- total de estudantes (*estudante*);
- população em idade ativa (*pia*);
- população masculina (*popmasc*);
- escolaridade média (*estudo*).

Já as variáveis em e_{ij}^f são:

- salário médio dos empregos formais (*salmed*);
- número de empresas (*empresa*);
- total de emprego formal nas empresas com quantidade de empregados entre 1 e 10 (*e10*);
- total de emprego formal nas empresas com quantidade de empregados entre 11 e 100 (*e100*);
- total de emprego formal nas empresas com quantidade de empregados entre 101 e 500 (*e500*);
- total de emprego formal nas empresas com quantidade de empregados entre 501 e 5000 (*e5000*);
- total de emprego formal nas empresas com quantidade de empregados acima de 5000 (*e10000*).

Buscando levar em consideração as variáveis com valor zero na estimação, optamos por trabalhar com as variáveis $e_{ij}^f + 1$ para calcular o logaritmo natural.

A partir da equação (1), fazendo regressões por setor SCNAE, podemos expandir o modelo para os anos de 1991 e 1996 através dos parâmetros $\hat{\beta}_{0i}$, $\hat{\beta}_{1ij}$ e $\hat{\beta}_{2ij}$ estimados para 2000. Desta forma, temos:

$$\ln \hat{e}_{ij}^t = \hat{\beta}_{0i} + \hat{\beta}_{1ij} \ln X_j + \hat{\beta}_{2ij} \ln e_{ij}^f, \quad (2)$$

onde $\ln \hat{e}_{ij}^t$ é a previsão do $\ln e_{ij}^t$ para cada ano e os demais parâmetros com o símbolo $\hat{}$ são os coeficientes estimados pela regressão de 2000.

Para cada ano, isolando o termo de emprego total, temos:

$$\tilde{e}_{ij}^t = \exp(\hat{\beta}_{0i} + \hat{\beta}_{1ij} \ln X_j + \hat{\beta}_{2ij} \ln e_{ij}^f), \quad (3)$$

onde $\tilde{e}_{ij}^t = \exp(\ln \hat{e}_{ij}^t)$.

Com isso, temos uma primeira estimativa do emprego total, a partir da formalidade, para 1991, 1996 e 2000, considerando o perfil de características locais e do mercado de trabalho de 2000 e a informação de onde cada trabalhador trabalha existente no Censo de 2000.

a.1) Calibragem

A partir desta primeira estimativa do emprego formal no nível município e setor SCNAE, fazemos uma calibragem para o nível unidade da federação, região metropolitana e setor SSCNAE. Para tanto, calibramos para o nível setor SSCNAE, unidade da federação e região metropolitana, de maneira que para cada *ufrmt* temos:

$$\sum_{mun=1}^m \sum_{SCNAE=1}^s e_{ij}^t = \alpha_{ij}^e \sum_{mun=1}^m \sum_{SCNAE=1}^s \tilde{e}_{ij}^t, \quad (4)$$

onde m é o número de municípios de cada *ufrmt* j , s é o número de setores SCNAE em cada SSCNAE i e α_{ij}^e é o ponderador do setor SSCNAE i e da *ufrmt* j .

Isolando este ponderador, podemos aplicá-lo sobre a previsão do emprego total, de maneira que:

$$e_{ij}^{tf} = \alpha_{ij}^e \tilde{e}_{ij}^t, \quad (5)$$

onde e_{ij}^{tf} é o emprego total final.

Entretanto, como α_{ij}^e é o ponderador, que aplicado sobre o exponencial da previsão, resulta no emprego total final, precisamos fazer uma transformação para transformá-lo em um ponderador, que aplicado sobre o emprego formal de determinada unidade (empresa, município ou unidade da federação), resulte no emprego total desta unidade. Assim, este novo ponderador será:

$$\gamma_{ijk}^e = e_{ij}^{tf} / e_{ijk}^f, \quad (6)$$

onde o índice k indica a unidade de referência.

Com isso, temos um ponderador que nos permite obter o emprego total a partir do emprego formal. Todavia, vale destacar que, no caso da unidade empresa, como a estimativa é feita no nível município e setor SCNAE, estamos considerando que todas as empresas do mesmo município e setor possuem o mesmo grau de informalidade. Apesar desta hipótese inicialmente não levar em consideração o porte da empresa, é importante lembrar que na estimação do modelo, controlamos pelo emprego formal de acordo com o porte da empresa. Tal controle visa considerar o fato de que as empresas de grande porte tendem a apresentar um elevado grau de formalidade, ao contrário de empresas de pequeno porte. Assim, este controle permite que o porte da empresa seja considerado em média. Contudo, em setores com grande discrepância com relação ao porte das empresas, esta hipótese homogeneiza as empresas de um mesmo setor.

b) Modelo de Massa Salarial

Para a massa salarial, o modelo proposto também utiliza as características dos municípios e do mercado de trabalho do ano de 2000 para estimar massa salarial total, pela equação:

$$\ln massa_{ij}^t = \delta_{0i} + \delta_{1ij} \ln X_j + \delta_{2ij} \ln massa_{ij}^f + \varepsilon_{ij}, \quad (7)$$

onde $massa_{ij}^t$ é a massa total (formal+informal) do setor i no município j , δ_{0i} é o termo constante da regressão de cada setor i , δ_{1ij} e δ_{2ij} são os coeficientes das variáveis explicativas da regressão de cada setor i , X_j é o conjunto de variáveis de características do município, $massa_{ij}^f$ é o conjunto de características do mercado de trabalho formal do setor i no município j e ε_{ij} é o resíduo da regressão.

As variáveis constantes em X_j são as mesmas do modelo de emprego. Já para as variáveis em $massa_{ij}^f$, optamos por trabalhar adicionando o valor 1 (um). Assim, consideramos as observações com valor zero na estimação. As variáveis constantes em $massa_{ij}^f$ são:

- massa salarial formal (*massa*);
- número de empresas (*empresa*).

Partindo das regressões por setor SCNAE da equação (7), podemos expandir o modelo para os anos de 1991 e 1996 através dos parâmetros $\hat{\delta}_{0i}$, $\hat{\delta}_{1ij}$ e $\hat{\delta}_{2ij}$ estimados para 2000. Assim, temos:

$$\ln \hat{massa}_{ij}^t = \hat{\delta}_{0i} + \hat{\delta}_{1ij} \ln X_j + \hat{\delta}_{2ij} \ln massa_{ij}^f, \quad (8)$$

onde $\ln \hat{massa}_{ij}^t$ é a previsão do $\ln massa_{ij}^t$ para cada ano e os demais parâmetros com o símbolo $\hat{}$ são os coeficientes estimados pela regressão de 2000.

Para cada ano, isolando o termo de emprego total, temos:

$$m\tilde{a}s\tilde{a}_{ij}^t = \exp(\hat{\delta}_{0i} + \hat{\delta}_{1ij} \ln X_j + \hat{\delta}_{2ij} \ln massa_{ij}^f), \quad (9)$$

onde $m\tilde{a}s\tilde{a}_{ij}^t = \exp(\ln \hat{massa}_{ij}^t)$.

Esta é uma primeira estimativa do emprego total para 1991, 1996 e 2000, considerando o perfil de características locais e do mercado de trabalho de 2000 e a informação de onde cada indivíduo trabalha, existente no Censo de 2000.

b.1) Calibragem

A calibragem deste modelo é exatamente igual à calibragem do modelo de emprego, partindo da expressão:

$$\sum_{mun=1}^m \sum_{SCNAE=1}^s massa_{ij}^t = \alpha_{ij}^m \sum_{mun=1}^m \sum_{SCNAE=1}^s m\tilde{a}s\tilde{a}_{ij}^t, \quad (10)$$

onde m é o número de municípios de cada *ufrmt* j , s é o número de setores SCNAE em cada SSCNAE i e α_{ij}^m é o ponderador do setor SSCNAE i e da *ufrmt* j .

Com este ponderador sobre a previsão do emprego total, temos:

$$massa_{ij}^{tf} = \alpha_{ij}^m \tilde{massa}_{ij}^t, \quad (11)$$

onde $massa_{ij}^{tf}$ é o emprego total final.

Transformando este ponderador em um ponderador, que aplicado sobre a massa salarial formal de determinada unidade (empresa, município ou unidade da federação), resulte na massa salarial total desta unidade, temos:

$$\gamma_{ijk}^m = massa_{ij}^{tf} / massa_{ijk}^f, \quad (12)$$

onde o índice k indica a unidade de referência.

Este ponderador que nos permite obter a massa salarial total a partir da massa salarial formal. Assim como no modelo de emprego, estamos considerando, no caso da unidade ser empresa, que todas as empresas do mesmo município e setor possuem o mesmo grau de informalidade.

Resultados

Por uma questão de praticidade na análise dos resultados, vamos apresentar os resultados consolidados, sem apresentar os resultados de cada uma das 112 regressões de setores SCNAE. No apêndice, apresentamos os coeficientes de explicação das regressões de cada um dos 112 setores SCNAE, lembrando que a unidade de observação é município. A partir desta abertura, na Tabela 1, mostramos o percentual de setores por faixa de coeficiente de explicação.

Tabela 1 – Percentual de Setores por Faixa de Coeficiente de Explicação

faixas	% setores	
	emprego	massa
abaixo de 0,5	6,25	10,71
entre 0,5 e 0,6	8,93	23,21
entre 0,6 e 0,7	18,75	37,50
entre 0,7 e 0,8	33,93	24,11
entre 0,8 e 0,9	29,46	1,79
acima de 0,9	2,68	2,68

Podemos observar que o modelo de emprego apresenta melhores resultados do que o de massa salarial. Além desta análise preliminar, podemos verificar a correlação entre o emprego total observado do Censo/PNAD (antes da calibragem) e o emprego total final, calculado considerando o ponderador de calibragem. O mesmo pode ser feito entre a massa salarial observada e a final, conforme a Tabela 2.

Tabela 2 – Correlação entre Variável Observada e Calculada (Modelo Proposto)

ano	Correlação	
	emprego	massa
1991	0,7949	0,8704
1996	0,9992	0,9943
2000	0,9856	0,9904

É importante destacar que o modelo de massa, apesar de ter menores coeficientes de explicação, apresenta uma correlação maior entre a variável observada e a calculada. Como os modelos utilizam regressões do ano de 2000 para expandir para os demais anos, é natural que a correlação de 2000 seja elevada. Além disso, o modelo de 1996 é no nível *ufrmt*, ao contrário dos de 1991 e 2000. Isto faz com que a correlação em 1996 seja próxima de 1 (um), uma vez que a calibragem também é no nível *ufrmt*, alterando apenas a agregação de setores. O principal resultado está justamente no ano de 1991, em que utiliza os resultados da regressão de 2000 e o modelo e a calibragem estão em níveis diferentes. Assim, podemos observar em 1991 uma alta correlação entre as variáveis analisadas, destacando a qualidade dos modelos propostos.

Anteriormente, em outro trabalho, propomos um modelo em que o emprego total e formal é agregado por *ufrmt* e SSCNAE. Em seguida, dividimos um pelo outro, resultando no ponderador que é aplicado diretamente sobre o emprego formal desagregado. Neste modelo, também podemos calcular a correlação entre emprego total final e emprego total observado, conforme Tabela 3.

Tabela 3 – Correlação entre Variável Observada e Calculada (Modelo Anterior)

ano	correlação
1991	0,4682
1996	1,0000
2000	0,7398

Comparando os resultados das tabelas 2 e 3, observamos que a correlação para o modelo proposto é superior a para o modelo anterior. Vale apenas ressaltar que, para 1996, o resultado da correlação é 1 (um) por construção, uma vez que a agregação inicial da PNAD já é *ufrmt* e usamos os setores SSCNAE ao longo de todo o modelo anterior.

Assim o modelo proposto neste trabalho permite a obtenção do emprego e massa salarial total (formal+informal) a partir do emprego e massa salarial formal, respectivamente, considerando a informação do local de trabalho do indivíduo. Apresentando uma melhoria significativa nos resultados com relação ao modelo anterior, o modelo proposto apresenta coeficientes de explicação das regressões por volta de 70% e elevada correlação entre a variável observada e a calculada em cada período analisado.

Aproximações para o preço médio

Nas tarefas de compatibilização de valor e tonelagem foram utilizadas informações de preços e de quantidade de duas grandes fontes: IBGE e SECEX. Em acréscimo, trabalhamos com informações adicionais obtidas por meio dos relatórios de análise de conjuntura setorial da consultoria LAFIS.

1. Informações do IBGE

As informações oriundas do sítio do IBGE foram relativas a dois segmentos: contas nacionais e produção e preços agrícolas. Para o primeiro caso trabalhamos com as informações sobre o PIB a preços de mercado para os anos de 2003 a 2005 e os respectivos consumos intermediários. Para o segundo caso trabalhamos com a PAM (Pesquisa Agrícola Municipal) para o ano de 2005.

a. Contas Nacionais

As informações relativas aos valores do PIB a preços de mercado são compostas da seguinte forma:

$$PIB_{pm} = PIB_{cf} + II - Sub.$$

Em que PIB_{pm} é o PIB a preços de mercado, II são os impostos indiretos e Sub os subsídios.

Além dos valores do PIB, que equivalem à soma do valor dos bens e serviços finais ou à soma dos valores adicionados em cada etapa da cadeia produtiva, trabalhamos com os valores intermediários, ou seja, a compra e venda de bens e serviços intermediários entre as empresas (dados do tipo Insumo-Produto para o ano de 2003) – ver quadro 1 abaixo.

Quadro 1 – Contas Nacionais

I - TABELA DE RECURSOS DE BENS E SERVIÇOS		
OFERTA	PRODUÇÃO	IMPORTAÇÃO
A	A₁	A₂
	=	+
II - TABELA DE USOS DE BENS E SERVIÇOS		
OFERTA	CONSUMO INTERMEDIÁRIO	DEMANDA FINAL
A	B₁	B₂
	=	+
COMPONENTES DO VALOR ADICIONADO		
C		

Fonte: IBGE

b. Pesquisa Agrícola Municipal

De acordo com o relatório metodológico do IBGE, a PAM é obtida por meio da rede de coleta do IBGE, através de consultas a entidades públicas e privadas, produtores, técnicos e órgãos relacionados à produção, comercialização, industrialização e fiscalização de produtos agrícolas. Utilizamos os informativos de 2003, 2004 e 2005.

Os dados da PAM originalmente contêm o valor total da produção de cada produto (são 64 produtos agrícolas) e a quantidade produzida (em peso) de cada produto. O preço por kg para cada produto foi obtido dividindo o valor total da produção pela quantidade produzida.

Para o ano de 2005, a PAM só tinha os dados dos produtos: algodão herbáceo, amendoim, arroz, aveia, centeio, cevada, feijão, girassol, mamona, milho, soja, sorgo granífero, trigo, triticale e algodão arbóreo.

Para os demais produtos ainda não havia informações para o ano de 2005. Para esses produtos foram usados os dados de 2004, sendo que os valores foram corrigidos pelo IPA-OG – grupo I: produtos agrícolas – FGV.

2. SECEX

As informações utilizadas a partir da SECEX foram obtidas por meio de extrações do sistema “Alice Web”, cuja base de dados é criada a partir das guias e formulários utilizados nas operações de comércio exterior junto ao SISCOMEX. Essas extrações permitem a obtenção de informações sobre tipo de produto, valor transacionado (FOB), quantidade (em kg), município de origem etc. O preço da exportação (ou da importação) que obtivemos foi pelo cálculo *Valor / Quantidade*, inicialmente expresso em US\$ FOB e convertido posteriormente em R\$ pela taxa média de câmbio do respectivo mês.

3. Relatórios Setoriais LAFIS

As informações de preço e quantidade construídas a partir dos relatórios setoriais da consultoria LAFIS são as que mais tiveram medidas discricionárias adotadas e são aquelas cujas fontes primárias são as mais diversas possíveis.

Os relatórios de conjuntura são elaborados por meio da composição de informações oficiais (órgãos de governo, institutos de pesquisa, entidades de classe etc.) e por estimativas e projeções elaboradas pela própria consultoria. As suas informações são de âmbito anual.

As informações mais sistematizadas referiam-se a *commodities*, cujos preços eram calculados por instituições (como a Esalq-USP no caso de vários produtos agrícolas) ou por meio das cotações internacionais (como no caso de minerais metálicos). Cotações em dólares foram convertidas e reais por meio da taxa anual média de câmbio disponibilizada pelo IPEADATA, cuja fonte primária é o Banco Central. Preços obtidos por meio de cotações internacionais não incorporam tributos locais.

Produtos manufaturados, por seu turno, apresentaram maiores problemas no cálculo do preço médio e da quantidade em kg ou toneladas. Para os setores que apresentavam informações de faturamento, o preço médio foi obtido pela divisão desse valor pela quantidade produzida.

No caso das medidas de quantidade, uma série de produtos é comercializada em unidades como arrobas, maços, centos, sacos de 50 kg, m2 etc. Essas medidas foram todas convertidas em peso, seja diretamente pela conversão de medidas afins, seja pela adoção de referências de peso, estes últimos expressos na tabela abaixo.

Tabela 1 – Medidas de conversão assumidas: de unidades de produto em kg do mesmo

	Peso médio (kg)
Laranja (unidade)	0,200
Maço Cigarros (unidade)	0,021

PC's (unidade)	25,000
Cerâmicas (m2)	14,667
Farmacêutico (unidade)	0,015

Ao final, obtivemos informações de preço em R\$ por kg e quantidade de produtos em kg. Destacamos que alguns “produtos” na verdade são setores, como tintas e vernizes, cujos dados de preço médio foram obtidos por faturamento do segmento produtor.

4. Calculando o preço médio por setor

Cada produto foi associado ao seu respectivo setor CNAE 5. Em seguida, o preço médio de cada setor CNAE 5 foi calculado. O preço médio do setor i é igual a média dos preços dos produtos (j) pertencentes ao setor (i) ponderados pela quantidade produzida (em peso) dos produtos.

$$p_i = \sum_{j \in i} \frac{Y_{j,i}}{\sum_{j \in i} Y_{j,i}} \cdot p_{j,i}$$

5. Calculando o preço médio por produto

Para os preços dos produtos (são 80 produtos – codificados pelo IBGE) usamos a mesma metodologia dos preços médios para os setores, descrita acima. Ou seja, se em um determinado setor CNAE 5 adotamos o preço da SECEX, os produtos associados a esse setor também terão a mesma origem de dados, i.e., SECEX. As exceções foram os preços da gasolina pura e do gasoálcool, que não conseguimos fazer nenhuma associação. Para estes dois produtos recorremos a uma nova fonte: ANP (Agência Nacional de Petróleo).

6. Metodologia adotada para cálculo dos preços médios setoriais finais

Comparamos os preços médios dos setores CNAE 5 da PAM, da LAFIS e da SECEX. A diferença de preços entre a base PAM e a base LAFIS era, em geral, pequena (nunca superior a 30%).

Quando a diferença de preços entre a SECEX e a PAM ou LAFIS foi superior a 30%, utilizamos o preço médio da base na qual a quantidade produzida (medida em peso) era superior. A idéia era que a base de informações com maior quantidade reportada nos daria a melhor representação da cesta de produtos do setor.

Quando a diferença de preços entre a SECEX e as outras bases foi inferior a 30%, adotamos o preço médio da SECEX.

O desvio padrão ponderado foi usado como um terceiro critério. Quando a diferença de preços era superior a 30% e as quantidades produzidas informadas nas bases eram próximas, o critério do menor desvio padrão ponderado foi utilizado na escolha da melhor fonte de informação.

Modelos Alternativos de Compatibilização

A primeira estratégia de implementação do sistema de compatibilização entre volumes físicos e financeiros partiu dos dados de valor adicionado dos municípios paulistas para o ano de 2005 da SEFAZ associados aos dados de Importação e Exportação por município da SECEX. Em princípios os dados pareciam bastante promissores. Era possível medir o valor adicionado enviado para fora do município mas dentro do Estado de São Paulo bem como o valor enviado para fora do Estado e para fora do país. Esses dados, fora os casos de sigilo, poderiam ser obtidos até o nível de 5 dígitos (mais de 500 setores). Combinado aos preços implícitos nos dados de importação e exportação, seria possível estimar o volume em toneladas saindo dos municípios em direção aos outros municípios do estado, a outros municípios do país e para fora do país.

No entanto, a primeira aproximação se demonstrou bastante incompatível com os poucos dados de carga e de preço que temos acesso com alguma segurança. A aproximação inicial se demonstrou incompatível com o total de carga movimentado dentro do estado estimado a partir da OD de 2005. Além do mais, o preço e o total de movimentação de alguns setores específicos se demonstrou incompatível com os relatórios setoriais realizados pela Dersa no âmbito do PDDT. A análise “anedótica” dos resultados indica que os problemas estão tanto nas estimativas de preço quanto nos dados de valor adicionado.

Ainda não temos segurança se será possível corrigir todas as disparidades. No entanto, foi possível ao menos qualifica-las um pouco mais. Originalmente estávamos trabalhando apenas com o valor adicionado. No entanto, não se transporta apenas o valor adicionado mas sim o valor total. Então, para compatibilizar volumes (físicos) com valores deve-se trabalhar com o valor total da produção. Isso não significa que não seja possível trabalhar com os dados de valor adicionado. Em princípio bastaria encontrar um multiplicador baseado na proporção de consumo intermediário do setor. Foi o que fizemos o que melhorou a aproximação. No entanto, ainda estamos distantes de uma compatibilização confiável entre valores monetários e físicos.

Reconhecendo os problemas existentes nas bases de dados propomos abaixo uma metodologia capaz de encontrar a melhor estimativa para preço e carga exportada por setor e por região. Como deve ficar claro, a melhor aproximação será para 7 setores e para as zonas agregadas da pesquisa OD para os Estado de São Paulo. A metodologia permite abrir os dados até a classificação CNAE de 5 dígitos e para municípios mas toda a calibragem será realizada para o Estado de São Paulo, na escala regional para 7 setores. O principal motivo para tal aproximação está ligado à precisão da pesquisa OD que permita utilizá-la como fator de calibragem como discutido a seguir.

Correções nas estimativas de preço

No que se refere aos preços, notamos que, via de regra, os preços parecem superestimados. Um dos motivos é certamente o que denominamos de “distorção cambial”. Os dados da SECEX disponíveis para essa análise referiam-se ao ano de 2003. Como os dados de Importação e Exportação estão em dólares, corrigimos os valores para real utilizando o dólar médio de cada mês levando em conta o mês da operação. Em seguida corrigimos os valores para 2005 utilizando o índice de inflação do setor. Ocorre que em 2003 o dólar ficou na faixa de 3 reais. Ao realizarmos essa operação partimos de um valor bem maior em reais que foi adicionalmente corrigido pela inflação. Esse certamente é um fator que está majorando os preços. Para corrigir

esse problema, o ideal seria utilizar os dados de Importação e Exportação para 2005 o que será realizado. No entanto, não temos segurança absoluta na base de 2005 obtida e por isso manteremos as tabulações de 2003 para servir como referência. Por outro lado, IPEADATA fornece uma série de índices de variação das importações e exportações divididos em alguns produtos que podem servir de base para a análise. Na ausência dessa informação no detalhamento necessário, utilizaremos os preços médios de 2003 em dólar corrigindo para 2005 simplesmente multiplicando-se pelo valor do dólar médio em 2005. As informações de variação de preço em dólar podem ser utilizadas para ajustar tal correção. Isso porque o sistema de correção proposto nesse parágrafo implica em preços constantes em dólares durante esses dois anos.

Ainda que o desvio padrão observado para os preços médios seja, via de regra, baixo, temos um problema adicional em função da definição de preço médio implicitamente adotada. O preço médio é encontrado dividindo-se o valor total de exportação e importação para cada setor identificável pelo peso total. Essa estratégia implica que o preço médio do setor é ponderado pela proporção do peso de cada produto ligado ao setor na pauta de exportações e importações. Mais formalmente:

$$p_i = \frac{\sum_p XM_{p,i}^{E,S}}{\sum_p W_{p,i}^{E,S}} = \frac{1}{\sum_p W_{p,i}^{E,S}} \sum_p \frac{W_{p,i}^{E,S} XM_{p,i}^{E,S}}{W_{p,i}^{E,S}} = \sum_p \frac{W_{p,i}^{E,S}}{\sum_p W_{p,i}^{E,S}} \frac{XM_{p,i}^{E,S}}{W_{p,i}^{E,S}} \quad (1)$$

Onde p_i é o preço médio (em dólares) do setor i ; $XM_{p,i}^{E,S}$ o valor total exportado ou importado (em dólares) para fora do país em uma determinada operação p (em qualquer município) com produtos ligados ao setor i estimada a partir dos dados da SECEX² e $W_{p,i}^{E,S}$ é o peso total dessa operação. Note que o nosso preço médio é uma soma do preço médio por quilo de cada operação, ponderada pela proporção do peso (em quilos) dessa operação no volume (físico) total de exportação e importação de todos os produtos ligados ao setor. Utilizar a proporção do peso como fator de ponderação é o mais adequado visto que o objetivo desse preço médio é permitir a transição de valores monetários para valores físicos. No entanto, a ponderação implica que a composição de cada produto (em termos de peso) na produção observada nas transações externas brasileiras deveria ser parecida com a composição observada nas transações internas. Não temos como garantir essa semelhança. Como o desvio padrão dos preços encontrados (estimado a partir da média ponderada no final do lado direito da equação acima) é consideravelmente baixo, ficamos de certa maneira confiantes nos dados de preço. No entanto, é óbvio que quanto maior a desagregação de preços melhor será a aproximação. Assim, o que devemos fazer é sempre trabalhar com a maior desagregação de preços possível. Para um setor mais agregado do que a máxima desagregação possível dos preços, o preço médio deve ser calculado como:

$$p_{j,m} = \sum_{i \in j} \frac{Y_{i,m} / p_i}{\sum_{i \in j} Y_{i,m} / p_i} p_i \quad i \in I^P, j \in I^I \quad \text{ou} \quad j \in I^3 \quad (2)$$

² Precisamos dos superescritos pois mais adiante devemos separar o total exportado para fora do país do total exportado para outras regiões do país. O segundo termo é necessário pois utilizaremos as estimativas da OD para corrigir problemas com a base da SECEX.

Onde $Y_{i,m}$ é o produto do setor i na região m , I^P é o universo (máximo) de setores para os quais foi possível calcular o preço médio enquanto I^I é o universo de setores compatível com a matriz de insumo-produto (42 setores – de agora em diante setor I-P) e I^3 é o universo de setores CNAE 3. Não temos uma classificação mais desagregada do que CNAE 5 então o universo máximo de preços é um subconjunto³ da CNAE 5. Precisamos dessa notação pois a seguir o universo de desagregação setorial passa a ser crítico. Assim, ponderamos o preço para uma determinada região a partir da sua estrutura de pesos e não a partir da estrutura das exportações. Se fosse possível uma desagregação muito elevada e assumindo uma economia aberta e perfeitamente competitiva, os preços internos deveriam ser iguais ao preço (FOB) de exportação ou de importação. Na prática sabemos que os mercados não são perfeitos o que poderia levar a algum erro também mas na verdade o nosso problema mais grave é obter dados de preço com precisão para uma desagregação elevada o suficiente⁴.

A dificuldade dessa definição de preço é que para CNAE 3 ou para os setores I-P não temos como recuperar o preço médio por setor diretamente a partir da base de preços. A conversão de valores para toneladas para essas duas desagregações precisa sempre ser realizada regionalmente. Quer dizer, a produção de um setor CNAE 3 ou I-P no município A será convertida para toneladas a um preço diferente do que no município B. Porém, a agregação de cargas (e não de valores depois convertidos para carga) a partir da classificação máxima (5 dígitos) deve dar o mesmo resultado visto que não temos como desagregar setorialmente acima desse nível. Assim, uma maneira alternativa para recuperar os totais físicos para as classificações mais agregadas será a partir da agregação dos dados de peso da CNAE 5, o que é totalmente equivalente a utilizar a aproximação (2) se considerarmos que $I^P \subset I^5$.

Com isso resolvemos parcialmente dois problemas graves que estão distorcendo a nossa medida de preço: a distorção cambial e a diferença entre a composição do comércio internacional e do comércio interno para níveis setoriais mais agregados. No entanto, mesmo para CNAE 5 (mais de 500 setores) alguns setores produzem centenas de produtos diferentes. Assim, se a composição de produtos do comércio internacional for muito diferente da composição interna para CNAE 5, os preços estimados a partir da SECEX permanecem distorcidos. A maneira como podemos identificar esse problema é computando o desvio padrão dos preços estimados a partir de (1). Na maioria dos casos encontramos um desvio bem baixo. Na realidade, ainda que o desvio estatisticamente correto para a estimativa proposta em (1) seja o desvio ponderado, o desvio sem ponderação fornece uma informação complementar. Imagine que a pauta de comércio internacional de um determinado setor esteja altamente concentrada em um produto. Ainda que esse setor compreenda produtos com preços bastante dispersos, a contribuição do produto predominante será muito mais significativa para a estimativa do desvio padrão (e do preço médio). Logo o desvio padrão não ponderado será bem maior do que o ponderado. Assim, será necessário estimar o desvio padrão ponderado e não ponderado. Além do mais, outra informação que precisa ser estimada é o total de operações observadas para cada produto. Todos esses dados fornecem dicas sobre preços mais problemáticos na nossa estimativa.

³ Apenas recuperamos preço para setores de produção de bens.

⁴ O problema torna-se mais grave pois trabalhamos com setores, que podem produzir bens bastante heterogêneos, e não com produtos, que seria mais fácil de se homogeneizar. A opção por uma divisão setorial permite que se utilize os dados de emprego com mais desagregação que são na prática os dados disponíveis nessa escala mais confiáveis que temos acesso.

Um segundo problema pode ocorrer se os preços de exportação e ou de importação estiverem sistematicamente acima ou abaixo de mercado interno. Isso pode ocorrer por diversos motivos. O preço FOB pode incluir taxas de intermediação que não existem no comércio interno. Não sabemos exatamente como estão distribuídos os impostos no preço FOB. O preço pode incluir mais impostos do que os preços internos no caso das importações e menos impostos no caso das exportações. Uma maneira de verificar possíveis distorções devido às assimetrias do sistema tributário portanto é observando as diferenças de preço na importação e na exportação. A outra fonte que utilizaremos para verificar possíveis diferenças entre os preços nos mercados interno e externos será a Pesquisa Agrícola Municipal (PAM). A partir dessa pesquisa é possível estimar preços para diversos produtos agrícolas. Comparando-se esses preços com os preços de exportação e de importação podemos ter uma visão de possíveis “**distorções residuais**”: distorções ligadas a outros fatores que não as distorções tributárias. Para ser possível comparar tais produtos, no entanto, é importante entender quais impostos estão (implicitamente ou explicitamente) incluídos no preço FOB de exportação, de importação e nos dados da PAM. Se as distorções residuais forem pequenas ou uniformes para os produtos agrícolas podemos ou ignorá-las para o mercado interno ou corrigir pelo fator encontrado no mercado agrícola, respectivamente. Adicionalmente a PAM pode servir para ajustar a composição interna dos produtos agrícolas caso existam agregações de produtos nos setores agrícolas da CNAE 5. Por esse motivo, a tabulação de preços da SECEX que permite comparação com a PAM deve seguir a classificação da PAM e não da CNAE 5. Logo, serão necessárias duas tabulações da SECEX.

Preços com desvio padrão (ponderado ou não) elevado devem ser estudados com mais cuidado. Observando a composição de produtos de cada setor com alto desvio padrão pode nos dar uma dica de como ajustar o preço para avaliar a produção física interna. No entanto, não temos fontes secundárias sistemáticas para criar uma regra de ajuste para todos os setores. O setor agrícola provavelmente será ajustado a partir da PAM como discutido no parágrafo anterior. Infelizmente os produtos agrícolas são justamente aqueles que apresentam menos heterogeneidade. O ajuste dos produtos industriais será realizado de maneira discricionária essencialmente utilizando as informações contidas nos relatórios setoriais. Em uma reunião entre Cepesp e Dersa devemos definir tais ajustes discricionários corrigindo “arbitrariamente” a matriz de preços. Pode-se fazer um segundo ajuste nesses preços em função da calibragem feita a partir dos dados da OD descrita a seguir.

Estimativa da produção, exportação e importação regional

Em princípio os dados do SEFAZ abririam uma possibilidade única. O dado de exportação e importação interna e externa de produtos teriam sido efetivamente medidos e não estimados como usualmente. Trata-se de uma base rara de se obter. Ocorre, no entanto, que a base apresenta diversas lacunas devido à sonegação e ao sigilo fiscal. Em primeiro lugar, mesmo quando observamos os totais para CNAE 3, notamos que o valor adicionado estimado para o Estado de São Paulo pelo SEFAZ é da ordem de 270 bilhões de reais enquanto o valor adicionado do estado estimado pelo Ibge é da ordem de 470 bilhões de reais. Além do mais, o total exportado por todos os municípios do estado para dentro do próprio estado é cerca de 50% mais elevado do que o total importado pelos municípios do estado de municípios do próprio estado (ou seja, a

importação é cerca de dois terços da exportação). Em princípio, em um sistema fechado⁵, todas as saídas deveriam corresponder a uma entrada.

Em poucos casos é possível recuperar o sigilo cotejando duas agregações diferentes. Em geral não há muito como corrigir o problema de sigilo. A sonegação poderia ser estimada seguindo procedimento semelhante ao utilizado para estimar o mercado de trabalho informal por setor. O problema é que não há nenhuma garantia que o grau de sonegação no mercado de trabalho seja semelhante à sonegação no mercado de produtos e não temos informação suficiente para recuperar os parâmetros para o produto. Além do mais, o fator de calibragem deveria ser o produto total por setor para o Brasil e não o total da massa salarial por setor e estado. Além do mais, os dados do SEFAZ não são replicáveis para fora do Estado de São Paulo nem para as zonas OD do metrô para a RMSP. Por esse motivo acabamos excluindo a possível alternativa de utilizar os dados da SEFAZ com fatores de correção. Nos pareceu arriscado demais para os nossos prazos. E, de todo modo, precisaríamos de uma outra alternativa para saber a precisão das estimativas.

A estratégia foi trabalhar com os dados de produção por setor e município estimado a partir dos dados de produção nacional por setor publicados pelo Ibge. O dado mais recente corresponde a 2003 e apresenta dados de valor adicionado e consumo intermediário para os setores I-P. Os dados de valor da produção do Ibge parecem coerentes com os alguns dados esparsos de produção física que tivemos acesso a partir dos relatórios setoriais o que aumenta a nossa confiança nessa base como ponto de calibragem. Para desagregar os dados de produção vamos simplesmente utilizar a proporção da massa salarial em cada região/subsetor, ou seja:

$$Y_{i,m} = \frac{S_{i,m}}{\sum_{i \in j} \sum_m S_{i,m}} Y_j \quad i \in I^3, j \in I^1 \quad \text{ou} \quad i \in I^5, j \in I^1 \quad (3)$$

Onde $S_{i,m}$ representa a massa salarial total (formal e informal) no setor i , região m . Isso será a nossa aproximação para o produto por setor por município em 2003. Y_j é o valor da produção do setor j no Brasil, ou seja, a soma do valor adicionado com o consumo intermediário para esse setor. Essa perspectiva é diferente da anterior. Primeiramente utilizamos o PIB municipal como medida básica e depois esse valor era dividido utilizando-se a proporção da massa salarial de cada setor na massa salarial total do município. Essa medida pode ter induzido a alguns erros pois o PIB municipal é também uma estimativa. Como comentamos acima, o produto setorial do país nos parece a aproximação mais confiável que temos e, portanto, a amarração se dará por essa variável.

O primeiro desafio consiste em corrigir esses valores para 2005. Em primeiro lugar, não está totalmente claro se devemos levar em conta o crescimento real ou nominal do produto. Se a nossa aproximação para preços em 2005 estiver de fato a preços de 2005 deveríamos utilizar a variação nominal. No entanto, se os preços estiverem em valor de 2003 apenas corrigindo a defasagem cambial, deveríamos utilizar a variação real. A estratégia será realizar as duas estimativas e definir a medida ideal na análise

⁵ O sistema é fechado por medida. A informação refere-se à saída (entrada) de produtos dos municípios do estado para (de) municípios do estado.

discricionária proposta para preços. Ou seja, a planilha básica que será criticada para se obter a melhor aproximação para preços conterá também os dados de produção. O ajuste para valores de 2005 pode ser ajustados discricionariamente. No IPEADATA mais uma vez, temos dados de variação (com algum grau de desagregação) do valor adicionado e do deflator implícito até 2005 (utilizar sempre as informações anuais). Essas duas informações devem ser utilizadas para termos uma medida de variação do PIB real e nominal para o Brasil. O índice será aplicado uniformemente para todas as regiões.

Uma vez obtido o valor total da produção por setor e por região precisamos obter uma estimativa para o valor das exportações de cada região para outras regiões dentro do país e para fora do país. Em princípio podemos utilizar os dados da SECEX para estimar as exportações de cada região para fora do país. Assumindo que o padrão permanece essencialmente o mesmo de 2003 para 2005 (ajustando para as variações nacionais) temos como realizar uma estimativa das exportações e importações para fora do país por região até a escala municipal. O único detalhe é que sabemos que os dados municipais da Secex superestimam as cidades portuária. Assim, seria interessante ajustar esses dados mas o ajuste depende de informações adicionais que podem não ser recuperáveis pela OD. No apêndice indicamos como realizar tal correção para os portos de São Sebastião e Santos se os dados do OD permitirem construir uma base de calibragem. Assumindo que temos dados confiáveis de exportação para fora do país (SECEX) e de produção por região (proposto em (3)) contruímos o nosso modelo. Parte do produto líquido dessas duas variáveis (produção menos exportação) será consumida internamente à região e outra parte será exportada para outras regiões do país. A questão é justamente separar essas duas parcelas. A estratégia adotada parte do conceito de Quociente de Localização utilizado em outras fases dessa pesquisa:

$$QL_{i,m} = \frac{S_{i,m} / \sum_i S_{i,m}}{\sum_m S_{i,m} / \sum_m \sum_i S_{i,m}} \quad (4)$$

Apenas para relembrar, o QL corresponde a uma medida de concentração relativa da atividade. Por exemplo, se a proporção (da massa salarial) do setor financeiro no município de São Paulo for 10% (o numerador da definição (4)) e a proporção no país for 5% (o denominador), o QL do setor financeiro em São Paulo será 2. Obviamente se o QL estiver acima de 1, a concentração (da massa salarial) está acima da média para essa região. A medida mais usual utiliza o emprego na definição do QL . Nesse caso utilizamos a massa salarial para corrigir para diferenças de produtividade partindo do princípio que o salário médio é uma aproximação para a produtividade do setor na região.⁶ Uma hipótese usual na literatura é que a parcela média representa a parcela consumida internamente enquanto a parcela acima da média deve ser exportada. Essa hipótese faz sentido se pensarmos que na média o Brasil consome 5% em serviços financeiros no exemplo acima. Cidades com 5% do seu produto no setor financeiro conseguem atender apenas à sua população. Nesse caso, a proporção da produção exportada ($x_{i,m}^1$) seria dada por:

⁶ Em equilíbrio perfeito a produtividade marginal do trabalho deve ser igual ao salário.

$$x_{i,m}^1 = \begin{cases} 0 & \text{se } QL_{i,m} \leq 1 \\ \frac{QL_{i,m} - 1}{QL_{i,m}} & \text{se } QL_{i,m} > 1 \end{cases} \quad (5)$$

Assumindo que as diferenças de produtividade entre as regiões podem ser captadas pelas diferenças de salário médio restam ainda 3 críticas principais a essa abordagem. Em primeiro lugar, implicitamente assume-se que não há comércio internacional. No exemplo anterior, pode ser que os brasileiros demandem apenas 4% do seu consumo para produtos do setor financeiro. No entanto, o Brasil exporta 1% da sua produção para outros países. Assim, uma cidade com 5% de sua produção em serviços financeiros poderia exportar 1% para outros lugares. Além do mais, assume-se (também implicitamente) que as regiões não importam e exportam produtos de um mesmo setor (ao mesmo tempo). Na prática sabemos que as regiões (pelo menos no que se refere aos países) em geral importam e exportam produtos de um mesmo setor.

Nos dois casos, o ponto de corte estabelecido para o QL (1) representa uma subestimação da parcela da produção exportada. Em apenas uma situação a exportação e a importação poderiam estar superestimadas. Se houver diferenças no padrão de consumo das regiões pode haver menos transações entre as regiões do que previsto pelo limite unitário estabelecido para o QL . Por exemplo, digamos que a demanda por ar condicionados no Nordeste seja maior do que no Sul. Uma cidade com QL 0,8 no setor de ar condicionados no Sul pode estar satisfazendo o mercado local assim como uma cidade com QL 1,2 no Nordeste. O limite unitário, no entanto, prevê que a cidade do Sul está importando ar condicionado e a cidade do Nordeste está exportando quando, na realidade, não está ocorrendo nenhuma transação dessas regiões com as demais nesse setor. O mais usual, no entanto, é que o limite unitário subestime o total exportado por uma região. Por esse motivo, adotamos alternativamente uma definição menos restritiva da proporção exportada nas regiões:

$$x_{i,m}^{rm} = \frac{QL_{i,m} - QL_{i,rm}}{QL_{i,m}} \quad (6)$$

Onde $QL_{i,rm}$ é o QL na região com o menor valor desse indicador (a região de requerimento mínimo). Por exemplo, se para algum setor houver uma região sem nenhuma empresa desse setor o requerimento mínimo será zero. Nesse caso qualquer região com ao menos uma empresa nesse setor exportará toda a sua produção. Isso ocorre, por exemplo, com o setor de petróleo. De fato, a proporção de exportação de uma refinaria é quase 100%. No entanto, essa medida pode ser um pouco exagerada para outros setores. Assim, via de regra, o limite de requerimento mínimo estará superestimando as exportações de uma região. Vamos usar as duas medidas como limites inferiores (limite unitário) e superiores (limite de requerimento mínimo) para a proporção do valor do produto exportado pelas regiões.

É fácil de perceber que o QL pode variar em função do grau de agregação setorial e regional. Conforme a agregação teremos resultados diferentes para o QL . Se a agregação fosse o Brasil todo, todos os QLs seriam iguais a 1 por definição. Assim, quanto menor a área de análise, maior será a disparidade de QLs e, portanto, maior será a transação entre as regiões (para qualquer definição). O mesmo vale para a

desagregação setorial onde utilizaremos uma classificação intermediária, a CNAE 3. Para a desagregação regional utilizaremos as zonas OD_STAN como região de análise também uma classificação intermediária entre os municípios as regiões funcionais⁷. Essa opção deve-se ao fato de que temos mais confiança nos dados de transporte de carga para essa agregação geográfica pois, dada a forma de coleta dos dados da OD, podem-se perder transações internas às zonas OD_STAN. Como utilizaremos os dados da OD para encontrar o ponto de corte para o *QL* intermediário entre o limite unitário e o limite de requerimento mínimo, é necessário utilizar os dados compatíveis para o cálculo do *QL*.

As equações (5) e (6) definem duas medidas de exportação total para cada região. Como temos uma medida da exportação da região para fora do país, podemos dividir a exportação máxima e mínima da região entre o seu componente destinado para outras regiões do país ($X_{i,m}^{I,rm}$ e $X_{i,m}^{I,l}$) e o destinado ao exterior ($X_{i,m}^{E,C}$). A exportação da região pode ser estimada diretamente pela SECEX preferencialmente com o ajuste proposto no apêndice. Essas variáveis são calculadas a partir da definição dos coeficientes de exportação acima, ou seja:

$$\begin{aligned} X_{i,m}^{I,rm} &= x_{i,m}^{rm} Y_{i,m} - X_{i,m}^E \\ X_{i,m}^{I,l} &= x_{i,m}^l Y_{i,m} - X_{i,m}^E \end{aligned} \quad (7)$$

A importação pode ser obtida por resíduo entre o consumo local e a produção local menos exportação (para fora do país e para outras regiões) mais as importações (de outras regiões ou países) mas reconhecendo o fato de que o total exportado internamente tem que ser igual ao total importado internamente. A estratégia será gerar uma necessidade de importação por região e depois distribuí-la de acordo com o total exportado em cada situação. Para chegarmos a uma medida de consumo regional, assumimos que a proporção da produção de produtos intermediários e finais de um setor permanece constante para as regiões e para divisões mais desagregadas do setor. Assim, classificando os produtos em bens intermediários e de consumo final podemos definir uma medida setorial que permite dividir a produção, exportação e importação nessas duas categorias:

$$c_i = \frac{1}{Y_i} \sum_{p \in in} Y_{p,i} \quad i \in I^I \quad (8)$$

Onde $Y_{p,i}$ é a produção do produto p pelo setor i ; $p \in in$ significa que estamos somando apenas para os produtos classificados como matéria prima ou intermediários; e, portanto, c_i representa a proporção de produtos intermediários produzidos pelo setor i . Como esse dado só está disponível para os setores I-P, temos que considerar $c_i = c_j$ onde $i \in I^3, j \in I^I, i \in j$. O primeiro passo é encontrar o consumo total de matéria-prima, bens intermediários e produtos finais para o Brasil. Isso será feito a partir de:

$$\begin{aligned} C_i^{in} &= c_i (Y_i - X_i^E + M_i^E) \\ C_i^f &= (1 - c_i) (Y_i - X_i^E + M_i^E) \end{aligned} \quad (9)$$

⁷ Vide relatório “Regiões Funcionais do Estado de São Paulo”.

Ou seja, o consumo interno de produtos do setor i é igual à produção do país (Y_i) menos a parcela exportada (X_i^E) mais a parcela importada (M_i^E) desse setor. Para dividirmos entre consumo intermediário (C_i^{in}) e consumo final (C_i^f) simplesmente utilizamos o coeficiente definido em (8). Para dividir os dois componentes de consumo entre as regiões vamos assumir que o consumo intermediário depende da massa salarial relativa ao país das empresas localizadas na região enquanto o consumo final depende da massa de renda relativa ao país dos residentes na região, ou mais formalmente:

$$C_{i,m} = C_{i,m}^{in} + C_{i,m}^f = \frac{\sum_i S_{i,m}}{\sum_i \sum_m S_{i,m}} C_i^{in} + \frac{R_m}{\sum_m R_m} C_i^f \quad (10)$$

Onde $C_{i,m}$ é o consumo total na região m que é composto pelos dois componentes de consumo regional: o consumo interno na região ($C_{i,m}^{in}$) e o consumo final ($C_{i,m}^f$). O consumo intermediário é dividido entre as regiões a partir da proporção da região na massa salarial ($S_{i,m}$) e o consumo final é dividido a partir da proporção da renda dos moradores da região m (R_m). A partir do consumo é possível encontrar um estimador da necessidade de importação dos municípios:

$$m_{i,m}^I + M_{i,m}^E = Y_{i,m} - X_{i,m}^I - X_{i,m}^E - C_{i,m} \quad (11)$$

Ou seja, aplicamos essencialmente mais uma identidade contábil. O consumo interno precisa ser satisfeito pela produção interna líquida de suas exportações mais as importações. Utilizamos o minúsculo para essa primeira aproximação das importações de produtos produzidos internamente pois precisamos ainda garantir que o total de importações internas seja igual ao total exportado. Ocorre que temos duas medidas de exportações internas: uma utilizando $QL = 1$ e outra utilizando o QL de requerimento mínimo como ponto de corte. Assim, precisamos incorporar a método de corte também na estimativa da importação interna:

$$\begin{aligned} m_{i,m}^{I,m} + M_{i,m}^E &= Y_{i,m} - X_{i,m}^{I,m} - X_{i,m}^E - C_{i,m} \\ m_{i,m}^{I,1} + M_{i,m}^E &= Y_{i,m} - X_{i,m}^{I,1} - X_{i,m}^E - C_{i,m} \end{aligned} \quad (12)$$

Como as exportações para fora do país e as importações de fora do país são medidas a partir dos dados da SECEX, a única diferença ocorre no indicador de exportação interna. No entanto, o critério de corte influencia a importação de outra maneira também. Ao garantir a igualdade entre exportação e importação interna, o total de comércio inter-regional estimado a partir do $QL = 1$ será bem menor do que o estimado a partir do requerimento mínimo. Isso fica claro no critério de distribuição definido abaixo:

$$M_{i,m}^{I,rm} = \frac{m_{i,m}^{I,rm}}{\sum_m m_{i,m}^{I,rm}} \sum_m X_{i,m}^{I,rm} \quad (13)$$

$$M_{i,m}^{I,1} = \frac{m_{i,m}^{I,1}}{\sum_m m_{i,m}^{I,rm}} \sum_m X_{i,m}^{I,1}$$

A partir de (7) e (13) temos duas medidas de exportação e duas de importação para os setores CNAE 3 utilizando dois critérios de corte distintos. A hipótese básica para chegar na proporção efetiva da produção na exportação de cada setor em cada município é justamente que o valor sempre estará entre os dois critérios. Lembre-se que os critérios representam dois limites (inferior e superior) para a parcela do valor da produção exportado para fora do país ou para outras regiões. Dada essa hipótese, o total exportado para outras regiões do país será dado por:

$$X_{i,m}^{I,*} = \alpha_i X_{i,m}^{I,rm} + (1 - \alpha_i) X_{i,m}^{I,1} \quad (14)$$

Onde $X_{i,m}^{I,*}$ é as estimativas “ótima” de exportação interna para o setor i na região m . A grande questão é como estimar o fator de distribuição entre os limites da proporção para cada setor (α_i). Como comentamos anteriormente, utilizaremos as estimativas agregadas da OD como ponto de calibragem. Para fazer o ajuste no valor “ótimo” entre esses dois limites assumimos que todas as outras medidas estão corretas. Em particular os preços médios já foram ajustados inclusive discricionariamente. Para compararmos os nossos dados com a OD no entanto, precisamos *i.* transformar nossos dados de valor em toneladas e *ii.* agregar os setores acima de tal forma que fiquem compatíveis com os setores da OD. A equação (9) faz exatamente esses dois ajustes:

$$X_{z,m}^{I,OD} = \sum_{i \in z} p_i^{-1} \alpha_i X_{i,m}^{I,rm} + p_i^{-1} (1 - \alpha_i) X_{i,m}^{I,1} \quad z \in I^{OD} \quad i \in I^3 \quad (15)$$

Onde $X_{z,m}^{I,OD}$ é a exportação em toneladas das empresas do setor z localizadas na região m para outras regiões do país estimado pela OD; e p_i^{-1} é o inverso do preço do setor calculado a partir de (2), ou seja, levando em consideração o preço para CNAE 5 na conversão. A informação no entanto não permite uma aproximação unívoca. Isso porque partimos de setores CNAE 3 com mais de 100 setores que precisam ser comparados com 7 setores. Em geral teremos pouco mais de 60 equações (as zonas OD_STAN do Estado de São Paulo) e cerca de 10 a 20 variáveis (o número de setores que compõem o setor mais agregado). Estamos no clássico problema de sobreidentificação. Como conhecido, a maneira de minimizar esse erro é encontrar os coeficientes de mínimos quadrados ordinários. Assim podemos encontrar o fator de distribuição entre os limites a partir da seguinte especificação:

$$X_{z,m}^{I,OD} - \sum_{i \in z} p_i^{-1} X_{i,m}^{I,1} = \sum_{i \in z} \alpha_i p_i^{-1} (X_{i,m}^{I,rm} - X_{i,m}^{I,1}) + \varepsilon_{z,m} \quad (16)$$

Onde $\varepsilon_{z,m}$ é um erro clássico com todas as característica usuais. Assim, podemos estimar α_i a partir de sete regressões onde a variável dependente é o lado esquerdo de (10)

acima e teremos tantas variáveis dependentes⁸ quantas forem necessários para completar o setor agregado a partir da CNAE 3 e forçamos que o constante seja zero. Aqui o sentido da regressão não é o usual, ou seja, não estamos usando como um modelo explicativo. O papel da regressão é encontrar a proporção da produção exportada pelos setores que minimiza a diferença entre os resultados estimados indiretamente e os medidos pela OD. Analogamente uma especificação para a importação interna seria dada por:

$$M_{z,m}^{I,OD} - \sum_{i \in z} p_{i,m}^{-1} M_{i,m}^1 = \sum_{i \in z} \alpha_i p_{i,m}^{-1} (M_{i,m}^{rm} - M_{i,m}^1) + \varepsilon_{z,m} \quad (17)$$

Onde as definições são totalmente análogas. Note-se que a estratégia implica em um pequeno número de observações dado a número de variáveis. Teremos sempre cerca de 130 observações (duas vezes o número de regiões OD_STAN no Estado de São Paulo – exportação e importação) para estimar em torno de 10 variáveis (os setores CNAE 3 que compõem o setor OD). Na realidade, se tivéssemos exatamente o mesmo número de regiões do que setores o problema seria unicamente identificável e teríamos apenas uma solução. Essa seria uma alternativa de análise que permitiria a consistência dos dados para alguma agregação regional do Estado de São Paulo. No entanto, nesse relatório estamos optando por um fator que minimize as diferenças regionais.

Além do mais, poderíamos estimar os α_i , apenas a partir dos dados de exportação e não empilhando as especificações (16) e (17). Isso faz sentido visto que os α_i , em última instância, estão definindo o QL de corte mais correto para o setor. O problema é que nesse caso, poderíamos ficar muito distantes das estimativas de importação corretas. Obviamente não faz sentido estimar parâmetros diferentes para a importação e para exportação pois, nesse caso, perderíamos a equivalência entre as medidas definida em (13). Dado os parâmetros podemos recuperar a proporção de exportações ótima como:

$$X_{i,m}^{I,*} = \alpha_i [x_{i,m}^{rm} Y_{i,m} - X_{i,m}^E] + (1 - \alpha_i) [x_{i,m}^1 Y_{i,m} - X_{i,m}^E] \quad (18)$$

Onde essencialmente substituímos (7) em (14). A equação (18) pode ser re-escrita como:

$$X_{i,m}^{I,*} = \alpha_i x_{i,m}^{rm} Y_{i,m} + (1 - \alpha_i) x_{i,m}^1 Y_{i,m} - X_{i,m}^E \quad (18')$$

Isso significa que o fator de correção ótima pode ser dado por:

$$x_{i,m}^{I,*} = \alpha_i x_{i,m}^{rm} + (1 - \alpha_i) x_{i,m}^1 \quad (19)$$

Substituindo (5) e (6) em (19) temos então que se $QL > 1$:

$$x_{i,m}^{I,*} = \alpha_i \frac{QL_{i,m} - QL_{i,m}^{rm}}{QL_{i,m}} + (1 - \alpha_i) \frac{QL_{i,m} - 1}{QL_{i,m}}$$

E se $QL \leq 1$:

⁸ $p_{i,m}^{-1} (X_{i,m}^{rm} - X_{i,m}^1)$

$$x_{i,m}^{I,*} = \alpha_i \frac{QL_{i,m} - QL_{i,rm}}{QL_{i,m}}$$

Combinando as duas temos então que:

$$x_{i,m}^{I,*} = \begin{cases} \alpha_i \frac{QL_{i,m} - QL_{i,rm}}{QL_{i,m}} & \text{se } QL_{i,m} \leq 1 \\ \frac{QL_{i,m} - [\alpha_i QL_{i,rm} + (1 - \alpha_i)]}{QL_{i,m}} & \text{se } QL_{i,m} > 1 \end{cases} \quad (20)$$

Com isso temos uma estimativa de exportação e importação de bens produzidos no país que, dados os parâmetros estimados como descrito acima, depende apenas da massa salarial por setor e por região, do total da renda por região, da produção do país por setor (e suas componentes), dos dados de exportação e importação com outros países por região e de um vetor nacional de preços. Todos esses dados são estimados a partir de fontes secundárias, ou seja, são exógenos ao nosso modelo.

As opções de desagregação como base para calibragem do modelo não implicam que não podemos mais recuperar dados mais desagregados setorialmente ou geograficamente. Os dados podem ser redistribuídos para CNAE 5 a partir da proporção da massa salarial como usualmente. A redistribuição para os municípios, no entanto, não pode ser realizada a partir de proporções. Isso porque não se pode desconsiderar as transações entre dois municípios da mesma região. Logo, é necessário recalcular a proporção exportada por cada município em função da concentração dos setores nessa escala. Porém o requerimento mínimo permanece o calculado a partir da desagregação em zonas OD_STAN. O fator de distribuição da proporção se altera pois, como podemos verificar em (20), o fator varia em função da região.

Essa metodologia permite então uma aproximação em teoria boa para a geração e atração de transportes da região além de fornecer o total de carga transportada (em toneladas) por setor dado o sistema de compatibilização calibrado dentro do modelo. Para generalizarmos esse modelo para a escala da Região Metropolitana, trabalhando com as zonas propostas pelo metrô, no entanto, não temos uma medida para todas as variáveis exógenas. Não sabemos as importações e exportações para fora do país em uma escala menor do que o município. Além disso, os dados da amostra do censo precisam ser substituídos pelos dados do universo. Isso significa que a variável de massa da renda total precisa ser substituída pela variável de massa da renda dos chefes de família. Essa segunda adaptação não nos parece problemática. Para lidar com a primeira restrição, a estimativa para as zonas da RMSP serão apresentadas somando-se o comércio interno e externo.