

# PAULO TADEU MEIRA E SILVA DE OLIVEIRA

Pessoas com deficiência: Análise  
crítica dos resultados do censo 2010 e  
novos desafios – Modelagem de  
Equações Estruturais

MAE/IME-USP  
SÃO PAULO 2014

# PAULO TADEU MEIRA E SILVA DE OLIVEIRA

Pessoas com deficiência: Análise  
crítica dos resultados do censo 2010 e  
novos desafios – Modelagem de  
Equações Estruturais

MAE/IME-USP  
SÃO PAULO 2014

# PAULO TADEU MEIRA E SILVA DE OLIVEIRA

Pessoas com deficiência: Análise  
crítica dos resultados do censo 2010 e  
novos desafios – Modelagem de  
Equações Estruturais

Projeto apresentado como parte das  
condições necessárias para obtenção de  
financiamento junto a entidades de  
fomento.

MAE/IME-USP  
SÃO PAULO 2014

## LISTA DE ILUSTRAÇÕES

FIGURA 1 – Distribuição de percentual de pessoas com pelo menos uma das deficiências investigadas na população residente dos municípios do Brasil em 2010.....	3
FIGURA 2 – Distribuição de percentual de pessoas com deficiência para: a) enxergar, b) ouvir, c) caminhar e d) intelectual.....	4
FIGURA 3 – Representação da CIF.....	11
FIGURA 4 – Modelo proposto para o projeto.....	34
FIGURA 5 – Algoritmo de procedimento de estimação para Modelo de equações Estruturais.....	52
FIGURA 6 – Óbitos por cólera em cada dia da epidemia.....	59
FIGURA 7 – Mapa de Londres onde deflagrou a epidemia de Cólera.....	60
FIGURA 8 – Médias e medianas para a escolaridade e renda na Ilha do Governador.....	63
FIGURA 9 – Matriz de proximidade espacial de primeira ordem normalizada pelas linhas.....	64
FIGURA 10 – Distribuição dos idosos na cidade de São Paulo (censo de 1991): a) distribuição dos valores por distribuição estatística, b) média móvel local.....	65
FIGURA 11 – IDH para São Paulo (censo de 1991).....	67

- FIGURA 12 – Variograma experimental do IDH para São Paulo (censo de 1991). Passo de amostragem 40km (tolerância 20km).....68
- FIGURA 13 – Diagrama de escalonamento de Moran para o índice de inclusão\exclusão social da cidade de São Paulo (censo de 1991),.....69
- FIGURA 14 – Mapa de espalhamento de Moran para o índice de inclusão\exclusão social da cidade de São Paulo (censo de 1991),.....70
- FIGURA 15 – Indicador de autocorrelação parcial para o índice de inclusão/exclusão da cidade de São Paulo (censo de 1991), para os valores com nível de significância maior que 95%.....71
- FIGURA 16 – Taxa total de mortalidade infantil por mil nascidos vivos no Rio de Janeiro em 1994.....72
- FIGURA 17 – Taxas de detecção média de hanseníase em menores de 15 anos, período1993-1997, por bairro de Recife, e, taxas estimadas através de alisamento bayesiano.75

## LISTA DE TABELAS

TABELA 1 – Decisões sobre evidência de Mod1 em relação a Mod2.....	34
---	----

## LISTA DE SIGLAS

AF – Análise Fatorial

AFC – Análise Fatorial Confirmatória

AGFI – Índice de Qualidade de Ajuste Corrigido

AIC – Critério de Informação de Akaike

BCC - Critério de Informação de Browne Cudeck

BIC - Critério de Informação Bayesiano

C – Índice de Geary

CFI – Índice Comparativo ajustado

CIF – Classificação Internacional de Funcionalidade,  
Incapacidade e Saúde

CDPD – Convenção dos Direitos das Pessoas com Deficiência

CONADE – Conselho Nacional dos Direitos da Pessoa Portadora  
de Deficiência

GFI – Índice de Qualidade de Ajuste

GLS – Mínimos Quadrados Generalizados

IBGE – Instituto Brasileiro de Geografia e Estatística

IDH – Índice de Desenvolvimento Humano

MEC – Ministério de educação

MEE – Modelagem de Equações Estruturais

MTE – Ministério do Trabalho e emprego

MV – Máxima Verossimilhança

I – Índice Global de Moran

$I_i$  – Índice do  $i$ -ésimo local de Moran

INSS – Instituto Nacional do Seguro Social

OIT – Organização Internacional do Trabalho

OMS – Organização Mundial de Saúde

RMR – Raiz do quadrado Médio Residual

RMSEA – Raiz do Erro Quadrático Médio

RV – Razão de verossimilhança

## SUMÁRIO

RESUMO.....	1
1 – INTRODUÇÃO.....	2
2 – SOBRE DEFICIÊNCIAS.....	11
3 – FINALIDADE E MOTIVAÇÃO.....	17
4 – OBJETIVOS.....	18
5 – PALAVRAS CHAVES.....	19
6 – JUSTIFICATIVAS PARA O PROJETO.....	20
7 – MATERIAIS E MÉTODOS.....	21
8 – DESCRIÇÃO DE ESTUDO.....	21
8.1 – DESCRIÇÃO DE VARIÁVEIS.....	21
9 – ANÁLISE EXPLORATÓRIA.....	23
9.1 – SELEÇÃO DE DADOS.....	23
9.2 – TESTE DE HOMOGENEIDADE.....	24
10 – AMOSTRAGEM.....	25
11 – IMPUTAÇÃO.....	27
12 – EQUAÇÕES ESTRUTURAIS.....	28
12.1 – CORRELAÇÕES ESPECIAIS.....	31
12.2 – TESTE DE HIPÓTESES.....	33
12.3 – ESTIMAÇÃO E AJUSTE.....	36
12.4 – SUPOSIÇÕES VIOLADAS E SUAS CONSEQUÊNCIAS.....	51
12.5 – MODELOS ELÍPTICOS.....	53
12.6 – SELEÇÕES DE VARIÁVEIS.....	54
12.7 – SELEÇÃO DE MODELOS.....	55
12.8 – SIMULAÇÕES.....	58
13 – GEOESTATÍSTICA.....	58
14 – RISCO DEFICIÊNCIA.....	77
14.1 – REGRESSÃO LOGÍSTICA BINÁRIA.....	79
14.1.1 – INTERPRETAÇÃO DOS COEFICIENTES DO MODELO.....	80
14.1.2 – REGRESSÃO LOGÍSTICA ORDINAL.....	81
14.1.3 – ESPECIFICAÇÃO DO MODELO ESTEREÓTIPO... ..	81
14.1.3.1 – INTERPRETAÇÃO DOS COEFICIENTES ESTIMADOS.....	82



14.1.3.2 – ESTIMAÇÃO DOS COEFICIENTES ESTIMADOS.....	83
14.1.3.3 – AVALIAÇÃO DA QUALIDADE DE AJUSTE.....	83
15 – CADASTRO PESSOA COM DEFICIÊNCIA.....	84
16 – FORMA DE ANÁLISE DE RESULTADOS.....	86
17 – CRIAÇÃO DO GRUPO DE PESQUISA ESTUDOS QUANTITATIVOS, QUALITATIVOS E COMPARATIVOS DA DISCAPACIDADE.....	86
18 – BENEFÍCIOS QUE PODEM SER OFERECIDOS POR ESTA PROPOSTA.....	87
19 – REFERÊNCIAS BIBLIOGRÁFICAS.....	88
20 – CRONOGRAMA DE ATIVIDADES.....	95
APÊNDICE A – CADASTRO DEFICIÊNCIA.....	95
APÊNDICE B – DESCRIÇÃO DAS VARIÁVEIS.....	110

## **PROPOSTA DE PROJETO E DE GRUPO DE PESQUISA**

Trata-se o projeto **Pessoas com deficiência: Análise crítica dos resultados do censo 2010 e novos desafios – Modelagem de Equações Estruturais** de um projeto de pesquisa que considera os diferentes tipos de deficiência (visual, auditiva, intelectual e física) em função de outras diversas variáveis relacionadas a identificação, instrução, família, trabalho, bens e moradia considerando o conjunto de dados obtidos das pessoas selecionadas para compor a amostra de respondentes que responderam o questionário completo do Censo Demográfico de 2010 do IBGE.

Com o intuito de substituir dados faltantes por valores plausíveis, utilizaremos técnicas de imputação, para que essa substituição possa ser feita de forma adequada considerando dados contínuos e categóricos.

Acontece que este trabalho vai muito mais além de uma proposta de projeto de pesquisa, por exemplo, o intervalo entre um censo e outro costuma ser de 10 anos e quando falamos de pessoas com deficiência nota-se demandas urgentes, contínuas e que necessitam de reparações mais imediatas, com dados sendo coletados de forma eficiente e com maior agilidade, eficiência, rapidez e que fiquem disponíveis em tempo real para que os pesquisadores e estudantes interessados possam efetuar suas análises e fazendo com que as condições de vida das pessoas com deficiência passem por processo de monitoração contínua e possibilitando outras questões como as relativas as condições de saúde e lazer.

Para este tópico, propomos a formação do cadastro de pessoas com deficiência de forma confidencial de modo que os respondentes não possam ser identificados, que a todo o momento possa ser atualizado e fiquem disponíveis para análise e monitoramento contínuo em tempo real

Para que possa melhor estudar as condições das pessoas com deficiência de acordo com sua localização e de entorno propomos uma modelagem Geoestatística e consideremos também a criação do índice risco deficiência, para que as condições de vida dessas pessoas sejam melhor monitoradas. Para esta parte consideramos a utilização de Análise Fatorial a partir da Modelagem de Equações Estruturais, Regressão Logística Binária, Regressão Logística Ordinal e Regressão Logística Estereótipo.

Entendemos também que existem outros projetos para serem propostos considerando, não apenas, o banco de dados do censo do IBGE, como também, outros conjuntos de dados originários de outras repartições como Ministério da Saúde, Educação, Trabalho e Emprego, Cultura, Transporte e outros de ordens federal, estadual e municipal e também de outros países, de modo que, as condições das pessoas com deficiência possam serem melhores estudadas e comparadas.

Diante deste cenário, nota-se a necessidade da criação de um grupo de pesquisa, que a princípio denominamos, ***Grupo de Estudos Quantitativos, Qualitativos e Comparativos sobre Discapacidade*** com característica multidisciplinar interagindo pesquisadores formados em áreas como Estatística, Ciências Sociais, Saúde, Psicologia e outros que tenham interesse em atuar nessa área de pesquisa.

# **Pessoas com deficiência: Análise crítica dos resultados do censo 2010 e novos desafios – Modelagem de Equações Estruturais**

Interessado: Dr Paulo Tadeu Meira e Silva de Oliveira

Unidade de Lotação: MAE – IME/USP

## **Resumo**

Para que haja um melhor suporte para as políticas públicas que visam uma melhor qualidade de vida a parcela da população formada por pessoas com deficiência, é de suma importância, antes de mais nada, descrever esse conjunto de pessoas sob o ponto de vista quantitativo, e como ponto de partida, começamos por estudar os dados do Censo Demográfico de 2010 realizado pelos recenseadores do Instituto Brasileiro de Geografia e Estatística (IBGE) do Questionário da Amostra (completo) formado por um conjunto de 20800804 entrevistados. Neste projeto, pretendemos iniciar com análises exploratórias investigando as diferentes deficiências com outras variáveis existente no banco de dados do censo; simulações para que possa avaliar o comportamento dos estimadores dos parâmetros e validar os modelos obtidos, a seguir, aplicação de um método de amostragem probabilística e complexa com o objetivo de selecionar uma sub amostra que seja representativa dos dados originais mantendo a sua proporcionalidade por região, estado e município dentre outras variáveis, com esses dados obtidos por amostragem e validar o modelo; tratamento por imputação de dados faltantes para que a matriz de dados seja completada para que possa dá continuidade ao trabalho com os outros procedimentos de análise de dados; modelagem de equações estruturais com o objetivo de avaliar relações existentes entre as variáveis observadas e as latentes, estudar a situação na qual é incorporada variáveis categóricas ordinais e nominais, seleção de variáveis em cada modelo para que em um mesmo modelo sejam incorporadas todas variáveis significantes e sejam excluídas as que não são significantes, implementar um método de seleção de modelos entre os diferentes modelos finais ajustados para cada uma das diferentes deficiências após a seleção das variáveis; com o intuito de avaliar a disposição espacial destes indivíduos aplicaremos métodos geoestatísticos e de estatística espacial; com o propósito de estabelecer um melhor monitoramento das condições de qualidade de vida do segmento pessoas com deficiência, propomos a criação do índice pessoa com deficiência; implementação de um sistema de cadastro on line para que o monitoramento da qualidade de vida das pessoas com deficiência possa ser feita em tempo contínuo; criação de um aplicativo on line com a finalidade

de obtenção de dados sobre pessoas com deficiência mediante cadastro para que essas análises possam ser feitas de forma mais rápida e ágil, sem ficar à espera exclusivamente dos dados do Censo como do IBGE que possui um intervalo de 10 anos para sua ocorrência; proposição e implementação de um grupo de pesquisa, que a princípio denominados de *Grupo de Estudos de Estudos Quantitativos, Qualitativos e Comparativos da Discapacidade*, como forma de integrar e relacionar com os diferentes pesquisadores trocando experiências e resultados nos trabalhos envolvidos e intercambio com outros grupos de pesquisas da área a nível nacional e internacional; e, por fim; avaliação crítica das variáveis obtidas nos questionários básicos e da amostra do censo 2010 para verificar a possibilidade de excluir variáveis existentes desnecessárias ou incluir outras variáveis que podem ser de interesse para melhorar a qualidade deste trabalho no próximo censo, ou, verificar a possibilidade desta parte dos dados serem incluídos no Questionário da população ou a possibilidade de existir um questionário em separado destinado a pessoas com deficiência. Em todos os casos pretendemos considerar conjunto de dados por estado, região e país.

## **Abstract**

### **Disabled people: Critical analyzes of results of the 2010 Census and new challenges – Structural Equations Model**

So that there is better support for public policies that seek a better life quality portion of the population consists of disabilities people, is of paramount importance, first of all, describe this group of people under the quantitative view point, and as a starting point, we begin by studying the data from the 2010 Demographic Census enumerators conducted by the Brazilian Institute of Geography and Statistics (IBGE) Sample Questionnaire (complete) formed by a set of 20,800,804 respondents. In this project, we intend to start with exploratory analyzes investigating different disabilities with other existing variables in the census database; simulations to help you evaluate the behavior of parameter estimators and validate the models obtained, then applying a method of probabilistic and complex sample with the aim of selecting a sub-sample that is representative of the original data while maintaining their proportionality by region state and county among other variables, with the data obtained by sampling and validate the model; treatment imputation for missing data so that the data matrix is completed so that you can continues to work with other data analysis procedures; structural equation modeling to evaluate the relationship between the observed variables and the latent, study the situation in which it is incorporated categorical ordinal and nominal variables, variavles selection in each model so that in one model all significant variables are incorporated and are excluded if they are not significant, implement a method of selecting among different models

adjusted for each different disabilities after the final variable selection models; in order to evaluate the spatial arrangement of these individuals will apply geostatistical and spatial statistical methods; in order to establish a better monitoring of life quality conditions of disabilities people segment, we propose the creation of the risk index disabilities persons; implementing an online registration system for monitoring the life quality of disabilities people can be made in continuous time; creating an application on line for the purpose of obtaining data on disabilities persons upon registering for these analyzes can be made more quickly and expeditiously, without waiting solely on data from the IBGE Census as having a range of 10 years for its occurrence; proposal and implementation of a research group, which first called the Study Group on Quantitative, Qualitative and Comparative Study of Disabilities, as a way to integrate and relate to different researchers sharing their experiences and results involved in the work and exchange with other groups research areas at national and international level; and, finally; critical evaluation of the variables obtained in the basic sample questionnaires and the 2010 census to verify the possibility of unnecessary or delete existing variables include other variables that may be of interest to improve the quality of this work in the next census, or to verify the possibility of this part of Questionnaire data were included in the population or the possibility of a separate questionnaire aimed at disabilities people. In all cases we intend to consider the data set by state, region and country.

## **1 - Introdução**

É, atualmente, considerado como fato, que sempre existiam ao longo da história pessoas com deficiência (Silva, 1986; Carvalho, 2001). Paulatinamente, as sociedades foram percebendo que, para além da caridade e da assistência, tais pessoas deveriam ser incluídas em programas e políticas públicas que pudessem valorizar seu potencial produtivo (Silva, 1986; Domingo, 2006; Figueira, 2008). Na realidade, as próprias pessoas com deficiência foram dando mostras de que podiam e desejavam estudar, trabalhar, terem suas vozes ouvidas e serem plenamente incluídas na sociedade (Garcia, 2010).

Segundo a Organização Mundial de Saúde (OMS) estima-se que mais de um bilhão de pessoas vivam com alguma forma de deficiência, algo próximo de 15% da população mundial (baseada em estimativas de 2010). Isto é menor do que as estimativas procedentes do IBGE, as quais datam também de 2010 e que sugerem aproximadamente 23,9% da população do Brasil, representada por 45606048 milhões, são pessoas com pelo menos uma deficiência. Dessas pessoas, 38473702 se encontram na área urbana e 7132347 em áreas rurais. A região nordeste

concentra os municípios com os maiores percentuais da população com pelo menos uma das deficiências investigadas, conforme ilustra a Figura 1, gráfico pertencente ao IBGE, que foi elaborado para o censo demográfico de 2010.

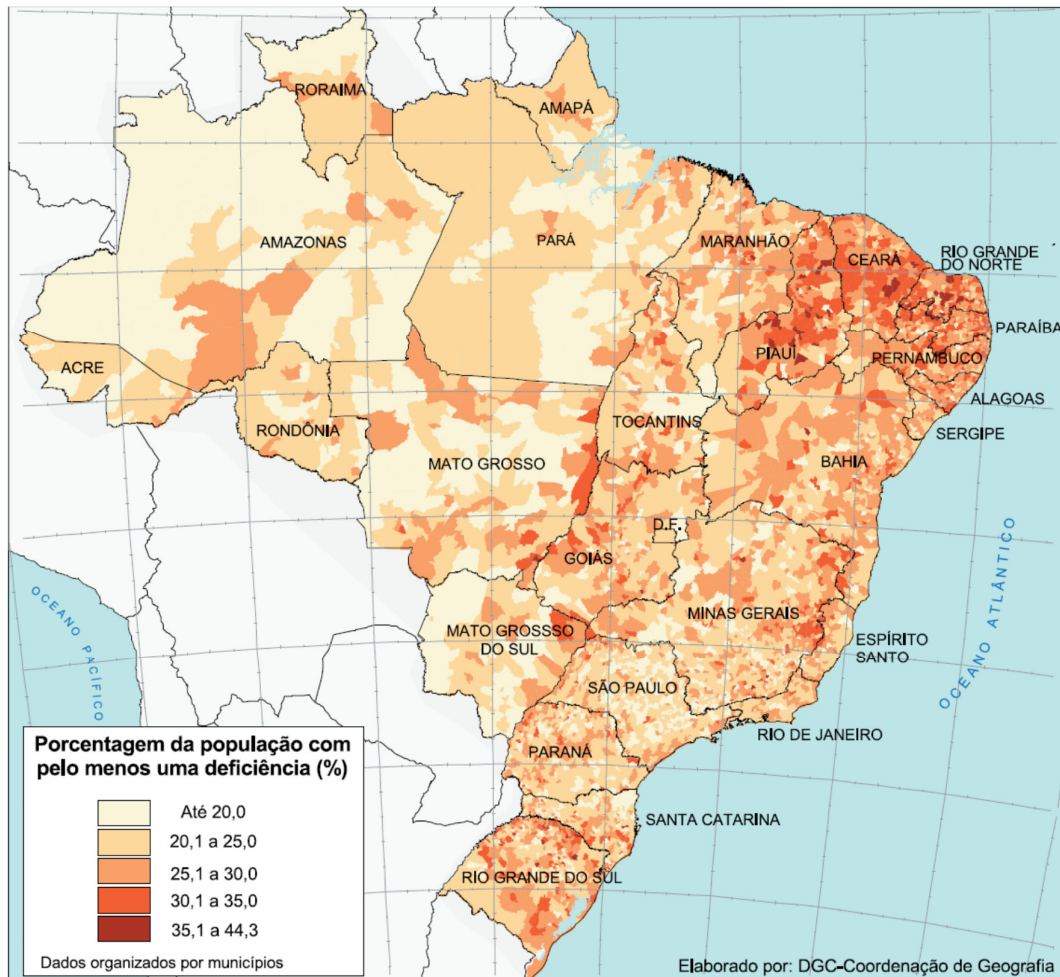


Figura 1. Distribuição de percentual de pessoas com pelo menos uma das deficiências investigadas na população residente dos municípios do Brasil em 2010.

A deficiência, continua a ser considerada um desafio universal com custos sociais e econômicos para indivíduos, famílias, comunidades e nações; varia de acordo com uma complexa combinação de fatores, incluindo idade, sexo, exposição a riscos ambientais, status sócios econômico, cultura e recursos disponíveis; estão associadas a problemas crônicos de saúde; envelhecimento global; e, finalmente; as pessoas com deficiência e os domicílios com algum membro deficiente enfrentam as piores realidades econômicas e sociais, se comparando às pessoas que não apresentam deficiência.

Observa-se no gráfico da Figura 1, que pessoas com deficiências estão mais concentradas na região nordeste e menos concentradas na região norte.

A seguir, mostraremos mapas de distribuição espacial elaborados pelo IBGE na Figura 2, de pessoas com deficiência para a) enxergar, b) ouvir, c) caminhar e d) intelectual.

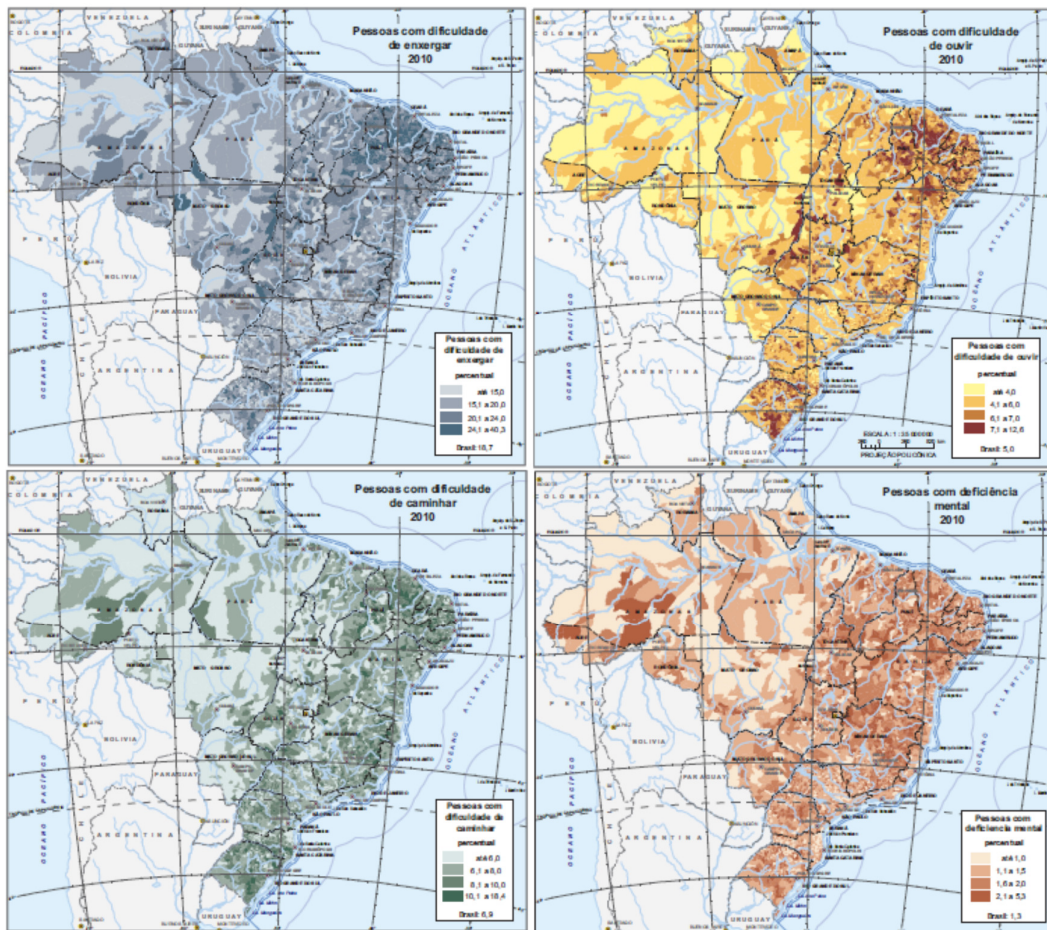


Figura 2. Distribuição do percentual de pessoas com deficiência para a) enxergar, b) ouvir, c) caminhar e d) intelectual.

Repare-se na Figura 2 que, existe uma maior quantidade de pessoas com deficiência visual, seguida de física, auditiva, e, por fim, intelectual.

Graças a mobilização dessas pessoas, foi possível garantir nos tempos atuais, um conjunto de leis que respalda esse contingente populacional, não só no que diz respeito ao mundo do trabalho, mas também quanto a direitos humanos como educação, saúde, lazer, e por fim; o direito de exercício pleno de cidadania.

Nos dias de hoje é fato reconhecido que a inclusão das pessoas com deficiência, tornou-se um assunto bastante discutido em termos de imprensa, políticas públicas, conjuntos de leis aprovados nas diferentes casas legislativas sancionadas pelos diferentes governos, e; muito discutido em diversos setores da sociedade entre outras (Santos, 2010).



Mesmo com esses avanços, sejam eles em termos históricos, como na mudança de paradigma, que permitem as pessoas com deficiência serem cidadãos, como no campo jurídico com existência de leis e decretos que estabeleceram esses direitos, porém, a participação das pessoas com deficiência, ainda, é muito restrita. Segundo dados do Censo Demográfico IBGE de 2010, estima-se que um pouco menos de 5% completaram o ensino superior, por outro lado, segundo o MEC (Ministério da Educação), esse número não chega a 2% e que um pouco mais de 30% das pessoas com deficiência em idade produtiva, estão no mercado formal segundo os dados do IBGE, mas, segundo os dados do MTE (Ministério do Trabalho e Emprego), esse número não chega a 5%, o que mostra, também, sérios problemas na consistência desses dados entre os diferentes setores do governo.

De qualquer forma, estes percentuais indicam que um número elevado de pessoas com deficiência continuam exercendo atividades informais, precárias e descontínuas ou simplesmente não possuem ocupação, vivendo com base em aposentadoria, pensões e/ou graças ao suporte familiar.

Acredita-se que as baixas condições de trabalho das pessoas com deficiência são devidos a situações como: dificuldade de acesso a educação, infraestrutura inadequada, preconceito, falta de informação e de melhores condições de acessibilidade por parte de escolas e empresas que fazem com que essas pessoas apresentem um menor nível de escolaridade o que dificulta o ingresso delas no mercado formal de trabalho.

Para a OMS, no que diz respeito a pessoas com deficiência nota-se: população de pessoas com deficiência em crescimento motivada por maior envelhecimento; experiências diversas resultantes da interação entre condições de saúde, fatores pessoais e ambientais; populações mais vulneráveis indicando prevalência maior em países de renda mais baixa; apresentam níveis de saúde mais precários; rendimento educacional inferior com taxas mais baixas de permanência e menor rendimento escolar; pior participação econômica com pessoas com deficiência mais propensas a ficarem desempregadas e ganhando menos quando empregadas; taxas mais altas de pobreza sofrendo altas taxas de privação com insegurança alimentar, habitação precária, falta de acesso à água limpa e ao saneamento e acesso inadequado a serviços de saúde, e, por fim; maior dependência e restrições a participação devido a confiança em soluções institucionais, falta de vida em comunidade e serviços inadequados.

A reversão desse quadro não é tarefa simples. Segundo especialistas (Garcia, 2010), isso exige ações em, pelo menos, onze aspectos centrais: a) a ampliação do conhecimento

público acerca das pessoas com deficiência e sua inserção na educação e no trabalho com melhora no acesso aos dados do IBGE e do MTE; b) as questões ligadas à legislação (não só em relação à chamada "Lei de Cotas", mas também à legislação trabalhista/previdenciária); c) o fortalecimento da inclusão escolar e das possibilidades de qualificação profissional, inclusive dentro das empresas como alternativa para o desenvolvimento pessoal da pessoa com deficiência rompendo situações de dependência que ainda existe e diminuir o déficit de quantidade de pessoas com deficiência que devem ser contratadas para atendimento da "Lei das Cotas" (Domingos, 2006; Garcia, 2010); d) Mais estímulos ao trinômio empreendedorismo no sentido de incentivar mais a livre iniciativa, inovação no sentido de estimular soluções mais criativas a um potencial mercado consumidor e funcionamento de pequenos negócios com noções de negócio, abertura e fechamento de empresa, noções de nichos de mercado e formas de captar os possíveis clientes como forma de oferecer um produto com alta probabilidade de ser bem recebido e ter boa saída por parte dos potenciais consumidores. Para isso, acreditamos que noções destes pontos devem fazer parte dos currículos escolares desde do nível fundamental para que essas ideias estejam mais amadurecidas nas cabeças dos estudantes nos níveis subsequentes e complementadas por parte do poder público e iniciativa privada com facilidades para abertura de pequenos negócios com custos mais subsidiados, financiamentos mais facilitados com juros menores oferecidos pelo governo e sistema financeiro e prazos maiores de pagamento e existência mais facilitada de suportes por entidades ligadas a pequenas e médias empresas como o SEBRAE e de inovação como da USP, UNICAMP e IPEN; e) acessibilidade como conceito síntese da sociedade inclusiva que muito mais que infraestrutura adequada, criar condições para que as pessoas com deficiência possam utilizar produtos, serviços e informações como qualquer outro cidadão; f) consolidação de novos paradigmas e formas de pensar a temática da deficiência, na sociedade em geral, mas especialmente entre as escolas (administradores, professores, funcionários e alunos) e empregadores (empresários ou gestores públicos) e as próprias pessoas com deficiência (Gasparetto et al., 2001); g) que fosse considerado obrigatório para os profissionais de saúde e de assistência social notificar ao Ministério da Saúde as pessoas atendidas e que tenham sido diagnosticadas com casos relacionados as deficiência com suas respectivas CIF's; h) estabelecimento de um sistema de cota de 5% das vagas para o ingresso das pessoas com deficiência nas escolas nos diferentes níveis educacionais como forma de estimular um maior número de pessoas com deficiência a adquirir ou melhorar o seu nível de instrução; i) implementar metodologias para a coleta de dados sobre as pessoas com deficiência, padroniza-los, deixá-los internacionalmente comparáveis para que possam estabelecer um ponto de referência e monitorar a evolução das políticas relacionadas; j) fortalecimento da

pesquisa sobre deficiência para que haja o aumento da compreensão pública, da oferta de informação para a elaboração de políticas dedicadas à deficiência, e, para a alocação eficiente de recursos, e, por fim; k) mais que Lei de Cotas, garantir que a pessoa com deficiência seja contratada para funções que correspondam ao seu verdadeiro potencial ou aptidão (Garcia, 2010; Rosa, 2009; Fontoura, 2006).

Além de tudo isso, é preciso que as condições econômicas e sociais do país evoluam positivamente. O crescimento econômico acelerado, uma melhor distribuição de renda, serviços públicos com qualidade e programas sociais eficazes, dentre outros aspectos, são benéficos para todos, inclusive, naturalmente, para aqueles com algum tipo de deficiência. Por mais que existam especificidades, não há um mundo "específico" das pessoas com deficiência. Elas também sentirão os efeitos da melhora social mais geral, por isso que as políticas específicas como gratuidades, cotas, isenções e benefícios não podem ser um fim em si mesmo, mas parte de uma estratégia mais ampla na qual, equiparando oportunidades, todos possam construir um país melhor, mais justo e humano.

Porém, para melhor avaliar a necessidade e o cumprimento desse arcabouço jurídico, torna-se necessário, melhor descrever esse conjunto de pessoas para saber respostas a questões como: Quantas são? Onde moram? Como vivem? Quais são as implicações que a deficiência acarreta no acesso dessas pessoas a todos os serviços humanos de forma autônoma e plena? Em suma, como a deficiência pode influenciar na qualidade de vida destas pessoas.

A deficiência é caracterizada como uma complexa experiência multidimensional, impõem inúmeros desafios de mensuração. As abordagens para mensurar a deficiência variam entre os diferentes países e influenciam os resultados. As medidas operacionais de deficiência variam de acordo com o objetivo e a aplicação dos dados, a concepção de deficiência, os aspectos da deficiência que se examina: deficiências, limitações para realizar determinadas atividades, restrições para participar de atividades, problemas de saúde relacionados, fatores ambientais – as definições, os tipos de questões levantadas, as fontes de informação, os métodos de coleta de dados e as expectativas de funcionamento.

Em termos estatísticos, mostra a existência de poucos estudos em termos formais, entre os quais destacam-se os dados obtidos através dos censos, que possibilita indagações como: Como as pessoas com deficiência estão distribuídas ao longo do país? Como avaliar o acesso das pessoas com deficiência em termos dos diferentes serviços mencionados anteriormente? Como está a evolução das pessoas com deficiência ao compará-las com as que

não apresentam deficiência? As diferentes deficiências são homogêneas? É possível formar grupos homogêneos? Quais seriam as variáveis que mais contribuem para os casos de deficiências? Existe um cadastro nacional das pessoas com deficiência? Como se encontram as pessoas com deficiência em comparação às que não apresentam deficiência? Resposta a essas e outras perguntas em termos estatísticos possivelmente poderão contribuir para melhor suporte a essas pessoas no sentido de serem melhores assistidas e de recursos serem melhor gerenciados e otimizados pelas ações das políticas públicas nesta área.

Já, segundo a OMS, em termos estatísticos, devemos melhorar as estatísticas nacionais sobre a incapacidade, melhorar a comparabilidade dos dados em nível nacional e internacional, e, por fim; desenvolver ferramentas apropriadas e preencher lacunas entre as diferentes pesquisas.

O Censo Demográfico, caracterizado como pesquisa por amostragem, é a mais complexa operação estatística realizada por um país, quando são investigadas as características de toda a população e dos domicílios do território nacional e constitui a mais importante fonte de referência para o conhecimento das condições de vida da população em todas as localidades.

Pesquisas por amostragem têm sido fontes de dados, tanto no nível individual quanto institucional, com objetivos tanto descritivos quanto analíticos desde o século XIX. Dados amostrais são a principal fonte de informações sócio demográficas da população, atividades econômicas, estilos de vida e opinião pública. Além disso, as tomadas de decisões em diferentes áreas do conhecimento são normalmente baseadas em grande extensão em dados amostrais (Cochran, 1977).

Nas últimas décadas, o rápido desenvolvimento das tecnologias de informação, tornou possível a geração, armazenamento e análise de grandes bases de dados. O uso de técnicas estatísticas, tanto exploratórias, como de modelagem, permite a obtenção de informações relevantes a partir destes dados, e estas informações servem como base para a tomada de decisões, políticas, administrativas e comerciais dentre outras.

No Censo Demográfico 2010, foram utilizados dois tipos de questionário: Questionário Básico que foi aplicado em todas as unidades domiciliares e Questionário da Amostra ou Completo que foi aplicado em todas as unidades domiciliares selecionadas para a amostra, que além da investigação contida no Questionário Básico, abrange outras características do domicílio e pesquisa importantes informações sociais, econômicas e demográficas dos seus moradores.

As questões relacionadas às pessoas com deficiências foram aplicadas no Questionário da Amostra que buscaram identificar as deficiências visual, auditiva e motora, de acordo com seu grau de severidade, através da percepção da população sobre essas suas dificuldades e para aquelas que declararam ter deficiência mental ou intelectual.

Neste trabalho aplicaremos técnicas de amostragem complexa probabilística com a finalidade de obter uma sub amostra da amostra original que satisfaça três requisitos básicos: menor tamanho amostral; que seja proporcional a níveis de variáveis como região, estado, município, deficiências para enxergar, ouvir, locomover e ouvir entre outras; e, por fim, que outras técnicas de análise como MEE (Modelagem de Equações Estruturais), AF (Análise Fatorial) e Regressão Logística possam ser executadas para a continuidade deste projeto.

Quanto aos dados obtidos no Censo Demográfico de 2010 permite inquirir o seguinte: Será que as questões que diz respeito a deficiência deveriam ser incluídas no questionário básico? Existem outras questões que devem ser incluídas mesmo no questionário completo? Deve existir um censo específico para pessoas com deficiência? Deveria existir um cadastro nacional de pessoas com deficiências? Deveriam existir outras questões a serem adicionadas? Respostas a essas e outras perguntas a este respeito pode auxiliar no melhor aprimoramento destas análises e desta pesquisa como um todo.

No processo de coletas de informações como no censo demográfico, algumas observações planejadas podem não ser obtidas, esses dados omissos ou faltantes (do inglês, *missings*) também requerem metodologia especializada para sua análise.

No caso deste projeto. Os dados são originários de entrevistas e problemas de preenchimento por parte dos agentes censitários do IBGE caracterizando uma causa típica de dados faltantes. Isso acontece por falta de conhecimento da resposta ou recusa intencional.

Pretendemos nesse trabalho adotar procedimentos de amostragem complexa para obter tamanho mínimo amostral que seja representativa e um banco de dados de amostra que possa ser melhor utilizada na continuidade deste projeto nas demais etapas e imputação de dados, para de alguma forma, substituir os dados faltantes por estimativas com o intuito de completar a tabela continuar a análise sem precisar descartar qualquer indivíduo participante da amostra.

Após conclusão da etapa de avaliação dos dados faltantes, para continuarmos este estudo consideraremos a MEE que permite o uso de relações separadas para cada conjunto de variáveis dependentes. O MEE nos fornece uma técnica de estimação apropriada e mais

eficiente quando se tem um conjunto de equações de regressão múltipla separadas, mas interdependentes, que devem ser estimadas simultaneamente. O MEE é caracterizado por dois componentes básicos: O modelo estrutural e o modelo de mensuração. O modelo estrutural é um conjunto de uma ou mais relações de dependência entre as variáveis latentes (construtos) do modelo; o modelo de mensuração especifica quais variáveis observadas (indicadores) que serão utilizadas como medida para cada variável latente, além de poder avaliar a confiabilidade desses indicadores para medir as variáveis latentes associadas. O MEE é resultante de uma evolução da modelagem de múltiplas equações desenvolvida principalmente nas ciências econométricas e originada dos princípios de mensuração da psicologia e das ciências sociais. Em termos simples, o Modelo de Equações Estruturais combina aspectos da regressão múltipla com análise fatorial para estimar uma série de relações de dependência inter-relacionadas simultaneamente (Gutierrez, 2005).

Nesta pesquisa, pretendemos calcular também medidas de ajuste como critério ICSF, critério ICS, Raiz do Erro Quadrático Médio (RMSEA), Índice de não centralidade de McDonald, índice de população Gamma, Índice de População Gamma Ajustado, Índice de Qualidade de Ajuste (GFI), Índice de Qualidade de Ajuste Corrigido (AGFI), Índice Comparativo Ajustado, Índice de Ajuste não normalizado ajustado, Índice de Ajuste Parcimonioso, Bollen's Rho, Bolle's Delta, Qui-quadrado Absoluto, Qui-quadrado Relativo, Critério de Informação de Akaike (AIC), Critério de Informação Bayesiano (BIC), Critério de Informação de Browne-Cudeck (BCC) e Raiz de Quadrado Médio Residual (RMR) para diferentes modelos obtidos segundo a modelagem de equações estruturais.

Considerando que estamos trabalhando com dados categóricos, em vista disso, propomos utilizar métodos de correlação como policórica, tetracórica, poliserial e biserial invés da correlação de Pearson que só é recomendável para dados contínuos.

Aplicaremos para seleção de variáveis dentro de cada modelo e entre os diferentes modelos utilizando critérios como AIC, BIC, Fator de Bayes, SSE, MSE, BCC, Cp,  $R^2$ , Estatística MV e RMR.

Os demais critérios serão utilizadas apenas para seleção de variáveis dentro de cada um dos modelos.

É uma técnica de ampla utilização e com diversas opções. As técnicas de MEE se distinguem pelas seguintes características: estimação de múltiplas e inter-relacionadas relações de dependência e a possibilidade de considerar variáveis observadas com erro de medição e variáveis não observadas ou variáveis latentes.

No caso deste projeto, pretendemos estudar as relações existentes entre as variáveis dependentes formada pelo bloco deficiência, variáveis independentes formada pelos blocos identificação, família, trabalho e instrução e as variáveis latentes (não observadas).

## 2 - Sobre as deficiências

Segundo a OMS, a deficiência é caracterizada como complexa, dinâmica, multidisciplinar e questionada. A transição de uma perspectiva individual e médica para uma perspectiva estrutural e social foi descrita como a mudança de um “modelo médico” para um “modelo social” no qual as pessoas são vistas como deficientes pela sociedade e não devido a seus corpos.

A Classificação Internacional de Funcionalidade, Incapacidade e Saúde (CIF) compreende funcionalidade e deficiência como uma interação dinâmica entre problemas de saúde e fatores ambientais. Promovido como um “modelo biológico psíquico-social”. A incapacidade é um termo abrangente para deficiências, limitações para realizar, e restrições para participar de certas atividades, que engloba os aspectos negativos da interação entre um indivíduo (com problema de saúde) e os fatores contextuais daquele indivíduo (fatores ambientais e pessoais) como pode ser verificado na Figura 3.

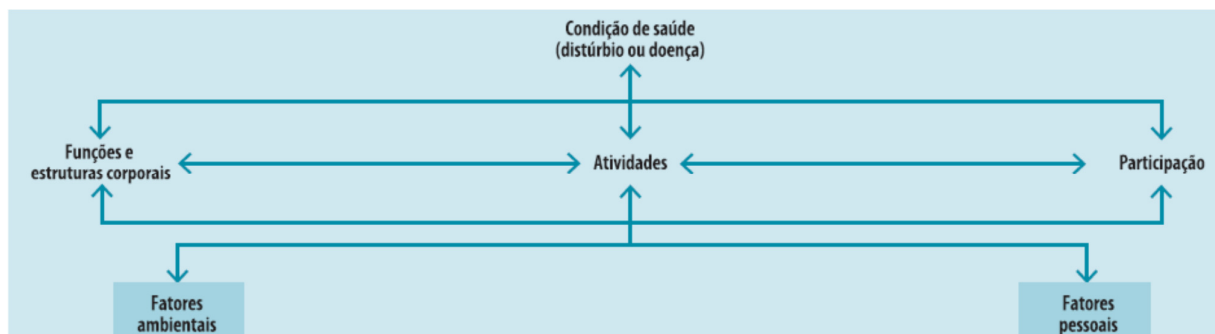


Figura 3. Representação da CIF

Repare-se na Figura 3 que a CIF é categorizada em alterações das estruturas e funções corporais, limitações e restrições à participação em certas atividades.

A CIF contém uma classificação de fatores ambientais que descreve o mundo no qual pessoas com diferentes níveis de funcionalidade devem viver e agir; reconhece fatores pessoais, tais como motivação e autoestima, que podem influenciar o quanto uma pessoa participa da vida em sociedade; universal por cobrir toda a funcionalidade humana e trata a deficiência como um contínuo; e, por fim; útil para uma ampla variedade de finalidades como pesquisa, fiscalização e informação.

A Convenção dos direitos das Pessoas com Deficiência (CDPD) reconhece que a deficiência é um conceito em evolução, mas realça também que “a deficiência resulta da interação entre pessoas com deficiência e barreiras comportamentais e ambientais que impedem sua participação plena e eficaz de forma igualitária”. Definir a deficiência como uma interação significa que a “deficiência” não é um atributo da pessoa. O progresso na melhoria da participação social pode ser realizado lidando com as barreiras que afetam pessoas com deficiência.

A deficiência é caracterizada por altas estimativas de prevalência, números crescentes, experiências diversas e populações vulneráveis.

Para as pessoas com deficiência, nota-se a existência de barreiras incapacitantes como políticas e padrões inadequados, atitudes negativas, falhas na oferta de serviços, problemas na proteção de serviços, financiamento inadequados, falta de acessibilidade, falta de consultas, envolvimento, falta de dados e evidência.

As vidas das pessoas com deficiência são afetadas por resultados de saúdes precários, rendimento educacional inferior, menor participação econômica, taxas mais altas de pobreza, maior dependência e restrição à participação e inclusão.

Para continuidade deste tópico, optamos por incluir informações sobre pessoas com deficiência, de como elas são caracterizadas em termo de legislação e também caracterizá-la de forma mais geral.

No Brasil há duas normas internacionais devidamente ratificadas, o que lhes confere status de leis nacionais, que são a Convenção nº 159/83 da OIT e a Convenção Interamericana para a Eliminação de Todas as Formas de Discriminação Contra as Pessoas Portadoras de Deficiência, também conhecida como Convenção da Guatemala, que foi promulgada pelo Decreto nº 3.956, de 8 de outubro de 2001. Ambas conceituam deficiência, para fins de proteção legal, como uma limitação física, mental, sensorial ou múltipla, que incapacite a pessoa para o exercício de atividades normais da vida e que, em razão dessa incapacitação, a pessoa tenha dificuldades de inserção social.

Nesse diapasão está o Decreto nº 3.298/99, cuja redação foi atualizada após longas discussões no Conselho Nacional dos Direitos da Pessoa Portadora de Deficiência (CONADE), pelo Decreto nº 5.926/04.

Pessoas reabilitadas, por sua vez, são aquelas que se submeteram a programas oficiais de recuperação da atividade laboral, perdida em decorrência de infortúnio. A que se atestar tal



condição por documentos públicos oficiais, expedidos pelo Instituto Nacional do Seguro Social (INSS) ou órgãos que exerçam função por ele delegada.

Veja-se, assim, o conteúdo da norma em comento: *Art. 3º Para os efeitos deste Decreto,*

*considera-se:*

***I - deficiência*** - *toda perda ou anormalidade de uma estrutura ou função psicológica, fisiológica ou anatômica que gere incapacidade para o desempenho de atividade, dentro do padrão considerado normal para o ser humano;*

***II - deficiência permanente*** - *aquela que ocorreu ou se estabilizou durante um período de tempo suficiente para não permitir recuperação ou ter probabilidade de que se altere, apesar de novos tratamentos; e*

***III - incapacidade*** - *uma redução efetiva e acentuada da capacidade de integração social, com necessidade de equipamentos, adaptações, meios ou recursos especiais para que a pessoa portadora de deficiência possa receber ou transmitir informações necessárias ao seu bem estar e ao desempenho de função ou atividade a ser exercida.*

#### **O que se entende por deficiência permanente?**

Considera-se deficiência permanente aquela que ocorreu ou se estabilizou durante um período de tempo suficiente para não permitir recuperação ou ter probabilidade de que se altere, apesar de novos tratamentos (Decreto nº 3.298/99, art. 3º, II).

#### **Incapacidade**

Considera-se como incapacidade uma redução efetiva e acentuada da capacidade de integração social, com necessidade de equipamentos, adaptações, meios ou recursos especiais para que a pessoa com deficiência possa receber ou transmitir informações necessárias ao seu bem estar e ao desempenho de função ou atividade a ser exercida (Decreto nº 3.298/99, art. 3º, III).

#### **O que é pessoa com deficiência habilitada?**

Aquela que concluiu curso de educação profissional de nível básico, técnico ou tecnológico, ou curso superior, com certificação ou diplomação expedida por instituição pública ou privada, legalmente credenciada pelo Ministério da Educação ou órgão equivalente, ou aquela com certificado de conclusão de processo de habilitação ou reabilitação profissional fornecido pelo INSS. Considera-se, também, pessoa portadora de deficiência habilitada aquela que, não tendo se submetido a processo de habilitação ou

reabilitação, esteja capacitada para o exercício da função (art. 36, §§ 2º e 3º, do Decreto nº 3.298/99).

### **O que é pessoa com deficiência reabilitada?**

Entende-se por reabilitada a pessoa que passou por processo orientado a possibilitar que adquira, a partir da identificação de suas potencialidades laborativas, o nível suficiente de desenvolvimento profissional para reingresso no mercado de trabalho e participação na vida comunitária (Decreto nº 3.298/99, art. 31).

A reabilitação torna a pessoa novamente capaz de desempenhar suas funções ou outras diferentes das que exercia, se estas forem adequadas e compatíveis com a sua limitação.

Por que se adota o termo pessoa portadora de deficiência ou pessoa com deficiência? A denominação utilizada para se referir às pessoas com alguma limitação física, mental ou sensorial assume várias formas ao longo dos anos. Utilizavam-se expressões como "inválidos", "incapazes", "excepcionais" e "pessoas deficientes", até que a Constituição de 1988, por influência do Movimento Internacional de Pessoas com Deficiência, incorporou a expressão "pessoa portadora de deficiência", que se aplica na legislação ordinária. Adota-se, hoje, também, a expressão "pessoas com necessidades especiais" ou "pessoa especial". Todas elas demonstram uma transformação de tratamento que vai da invalidez e incapacidade à tentativa de nominar a característica peculiar da pessoa, sem estigmatizá-la. A expressão "pessoa com necessidades especiais" é um gênero que contém as pessoas com deficiência, mas também acolhe os idosos, as gestantes, enfim, qualquer situação que implique tratamento diferenciado. Igualmente se abandona a expressão "pessoa portadora de deficiência" com uma concordância em nível internacional, visto que as deficiências não se portam, estão com a pessoa ou na pessoa, o que tem sido motivo para que se use, mais recentemente, conforme se fez ao longo de todo este texto, a forma "pessoa com deficiência". Esta é a denominação internacionalmente mais frequente, conforme demonstra Romeu Kazumi Sasaki.

### **Tipos de Deficiência**

#### **Deficiência física:**

É a alteração completa ou parcial de um ou mais segmentos do corpo humano, acarretando o comprometimento da função física, apresentando-se sob a forma de paraplegia, paraparesia, monoplegia, monoparesia, tetraplegia, tetraparesia, triplegia, triparesia, hemiplegia, hemiparesia, ostomia, amputação ou ausência de membro, paralisia cerebral,

nanismo, membros com deformidade congênita ou adquirida, exceto as deformidades estéticas e as que não produzam dificuldades para o desempenho de funções (Decreto nº 5.296/04, art. 5º, §1º, I, "a", c/c Decreto nº 3.298/99, art. 4º, I).

Para melhor entendimento, seguem-se algumas definições:

**Amputação** - perda total ou parcial de um determinado membro ou segmento de membro;

**Paraplegia** - perda total das funções motoras dos membros inferiores;

**Paraparesia** - perda parcial das funções motoras dos membros inferiores;

**Monoplegia** - perda total das funções motoras de um só membro (inferior ou superior);

**Monoparesia** - perda parcial das funções motoras de um só membro (inferior ou superior);

**Tetraplegia** - perda total das funções motoras dos membros inferiores e superiores;

**Tetraparesia** - perda parcial das funções motoras dos membros inferiores e superiores;

**Triplegia** - perda total das funções motoras em três membros;

**Triparesia** - perda parcial das funções motoras em três membros;

**Hemiplegia** - perda total das funções motoras de um hemisfério do corpo (direito ou esquerdo);

**Hemiparesia** - perda parcial das funções motoras de um hemisfério do corpo (direito ou esquerdo);

**Ostomia** - intervenção cirúrgica que cria um ostoma (abertura, ostio) na parede abdominal para adaptação de bolsa de fezes e/ou urina; processo cirúrgico que visa à construção de um caminho alternativo e novo na eliminação de fezes e urina para o exterior do corpo humano (colostomia: ostoma intestinal; urostomia: desvio urinário);

**Paralisia Cerebral** - lesão de uma ou mais áreas do sistema nervoso central, tendo como consequência alterações psicomotoras, podendo ou não causar deficiência mental;

**Nanismo** - deficiência acentuada no crescimento. É importante ter em mente que o conceito de deficiência inclui a incapacidade relativa, parcial ou total, para o desempenho da atividade dentro do padrão considerado normal para o ser humano. Esclarecemos que a pessoa com deficiência pode desenvolver atividades laborais desde que tenha condições e apoios adequados às suas características.

**Deficiência auditiva:**

É a perda bilateral, parcial ou total, de 41 decibéis (dB) ou mais, aferida por audiograma nas frequências de 500Hz, 1.000Hz, 2.000Hz e 3.000Hz (Decreto nº 5.296/04, art. 5º, §1º, I, "b", c/c Decreto nº 5.298/99, art. 4º, II).

**Deficiência visual:**

De acordo com o Decreto nº 3.298/99 e o Decreto nº 5.296/04, conceitua-se como deficiência visual:

**Cegueira** - na qual a acuidade visual é igual ou menor que 0,05 no melhor olho, com a melhor correção óptica;

**Baixa Visão** - significa acuidade visual entre 0,3 e 0,05 no melhor olho, com a melhor correção óptica;

Os casos nos quais a somatória da medida do campo visual em ambos os olhos for igual ou menor que 60º; Ou a ocorrência simultânea de quaisquer das condições anteriores. Ressaltamos a inclusão das pessoas com baixa visão a partir da edição do Decreto nº 5.296/04. As pessoas com baixa visão são aquelas que, mesmo usando óculos comuns, lentes de contato, ou implantes de lentes intraoculares, não conseguem ter uma visão nítida. As pessoas com baixa visão podem ter sensibilidade ao contraste, percepção das cores e intolerância à luminosidade, dependendo da patologia causadora da perda visual.

**Deficiência mental:**

De acordo com o Decreto nº 3.298/99, alterado pelo Decreto nº 5.296/04, conceitua-se como deficiência mental o funcionamento intelectual significativamente inferior à média, com manifestação antes dos 18 anos e limitações associadas a duas ou mais áreas de habilidades adaptativas, tais como: comunicação; cuidado pessoal; habilidades sociais; utilização dos recursos da comunidade; saúde e segurança; habilidades acadêmicas; lazer; e trabalho. (Decreto nº 5.296/04, art. 5º, §1º, I, "d"; e Decreto nº 3.298/99, art. 4º, I).

**Deficiência múltipla:**

De acordo com o Decreto nº 3.298/99, conceitua-se como deficiência múltipla a associação de duas ou mais deficiências.

A condição de pessoa com deficiência pode ser comprovada por meio de:

Laudo médico, que pode ser emitido por médico do trabalho da empresa ou outro médico, atestando enquadramento legal do(a) empregado(a) para integrar a cota, de acordo com as definições estabelecidas na Convenção nº 159 da OIT, Parte I, art. 1; Decreto nº 3.298/99, arts. 3º e 4º, com as alterações dadas pelo art. 70 do Decreto nº 5.296/04. O laudo deverá especificar o tipo de deficiência e ter autorização expressa do(a) empregado(a) para utilização do mesmo pela empresa, tornando pública a sua condição;

Certificado de Reabilitação Profissional emitido pelo INSS.

### **3 – Finalidade e motivação**

A finalidade deste projeto é aplicar procedimentos estatísticos estudados nesta proposta, ao conjunto de dados do Censo Demográfico de 2010 do IBGE para o país, por região, estado e município. Este projeto propõe, também, o estudo, desenvolvimento e aplicação de técnicas de análise e modelagem adequadas a banco de dados do tipo censo demográfico, e mais especificamente, investigação das diferentes deficiências estudadas com outras variáveis coletadas no censo como nível de instrução e faixa salarial em salários mínimos utilizando procedimentos de seleção de variáveis de modo que para cada uma das diferentes deficiências tenha um modelo ajustado com as variáveis que sejam significativas.

No que diz respeito a seleção entre os diferentes modelos, a dificuldade nestes é a comparação de modelos que não necessariamente são encaixados (sequência de modelos, cujos modelos reduzidos com menor quantidade de parâmetros, são subconjuntos dos modelos saturados que apresentam maior quantidade de parâmetros).

Vários procedimentos de seleção de modelos e critérios estatísticos podem ser considerados para auxiliar na elaboração de decisões sobre os termos que devem ser incluídos ou excluídos do modelo. Em geral, a seleção de modelos está ligada à estimação dos parâmetros associados ao modelo estatístico adotado ou à predição do comportamento de variáveis aleatórias envolvidas na definição do modelo (Burnham and Anderson, 1998, 2002). Metodologias, como SSE, AIC e BIC, são bastante flexíveis (Blangero et al., 2001) e podem ser utilizadas com vantagem se acopladas a algoritmos computacionais, como modelagem de equações estruturais (Oliveira, 2008).

Os principais problemas neste estudo são: grande tamanho amostral, dados faltantes devido ao não preenchimento por desconhecimento ou omissão intencional por parte dos entrevistados, o que dificulta em termos de análise de dados exigindo estudo de pesquisa por amostragem e máquinas mais potentes para esse tipo de análise; os dados são categóricos que ao optar pela utilização da MEE necessita um estudo mais detalhado de como montar a matriz de

correlação considerando dados categóricos ordinais e não ordinais, existência de diversas suposições que são violadas necessitando de adoção de procedimentos de correção e como fica os resultados das medidas de estimação e qualidade de ajuste para este tipo de dados, pois, a técnica de MEE sempre foi recomendada para dados contínuos (Bistaffa, 2010).

Um outro ponto importante a ser discutido são as suposições do modelo que são violadas quando consideramos variáveis categóricas ordinais e nominais ao invés de variáveis contínuas e suas consequências na estimação do modelo e na validade das estimativas obtidas. Uma possível solução para corrigir as suposições violadas é apresentar procedimentos necessários para corrigir para que possa incorporar variáveis categóricas na MEE.

Uma outra alternativa, neste caso, quando a suposição de normal multivariada para o vetor de variáveis observadas é violada, é utilizar outras distribuições com cauda mais leves ou mais pesadas do que as da normal. Classe esta de distribuições conhecida como elíptica (Melhado, 2009).

Índice risco deficiência para que possa servir de suporte e monitoramento das políticas públicas direcionadas à pessoas com deficiência e seus familiares.

Cadastro deficiência para que possa permitir maior agilidade, não ficar dependente dos dados obtidos exclusivamente dos trabalhos tipo censo, permissão de análise em tempo real e melhor monitoramento por parte dos pesquisadores interessados, conforme modelo a ser desenvolvido contendo questões sobre acessibilidade, infraestrutura nas mais diversas repartições públicas e privadas, acessibilidade na infraestrutura urbana, além dos diferentes tópicos descritos na seção 8.3 descrição das variáveis, tendo como potenciais respondentes pessoas com deficiência e que não apresentam deficiência, o que permite estudos comparativos para estudos e modelagens.

Aplicação de técnicas geoestatísticas com o intuito de avaliar a situação de infraestrutura de entorno e das residências das pessoas com deficiência e de seus familiares.

Criação do grupo de pesquisa de estudos quantitativos, qualitativos e comparativos sobre incapacidade, com o intuito de interagir com outros pesquisadores, maior alavancagem com outros projetos sendo desenvolvidos simultaneamente e formação de novos pesquisadores que possibilite que esse grupo prospere e não sofra descontinuidade.

#### **4 - Objetivos**

O projeto tem como objetivo descrever o conjunto de pessoas com deficiência conforme definido no Censo Demográfico de 2010 elaborado pelo IBGE divididos em problemas para enxergar, ouvir, locomover e mental e estudar variáveis que podem servir de comparação entre o conjunto de indivíduos que apresentam pelo menos uma das diferentes deficiências mencionadas acima e o conjunto de indivíduos que não apresentam deficiência.

Implementar procedimentos de imputação para que a matriz seja completada para outros procedimentos de análises, de modo, que não seja necessário descartar dados de pessoas entrevistadas por ocorrência de dados faltantes.

Identificar a origem interna e externa da informação e extrair o subconjunto de dados necessários (selecionar variáveis de interesse) para a aplicação das outras análises.

Determinar o tamanho mínimo necessário e selecionar os indivíduos que comporão a nova amostra a ser utilizada na análise por MEE, em termos amostrais, e, de modo que garanta que a matriz de dados estejam completa para a aplicação de outros métodos de análise como modelagem por equações estruturais.

Descobrir modelos que tenham sentido em termos teóricos e que apresente correspondência estatística razoável com os dados coletados, isto é, no caso deste projeto, que melhor explique o relacionamento entre as variáveis do bloco deficiência e dos blocos identificação, família, trabalho e instrução.

Descrever modelos que apresentam variáveis explicativas que mais influenciam as diferentes deficiências em função de outras variáveis como relacionadas a trabalho, nível de instrução e de identificação dos indivíduos.

Reorganizar a distribuição das variáveis explicativas a fim de detectar as mais significativas para explicar semelhanças ou dissimilaridades entre elas.

Estudar o padrão de associação entre as diferentes variáveis independentes para cada uma das diferentes deficiências.

Criação e implementação do índice risco pessoa com deficiência, que dado conhecimento de sua incidência, pode ser utilizado como instrumento facilitador da avaliação das condições de vida das pessoas com deficiência quanto a acesso a direitos humanos como educação, trabalho, lazer, moradia, transporte, comunicações e saúde.

Geoestatística, com o intuito de avaliar a infraestrutura formada pelo entorno das localizações das residências das pessoas com deficiência e de seus familiares.

Criação de um cadastro contendo dados sobre pessoas com deficiência para que possa permitir análise e emissão de resultados de forma mais rápida e mais ágil para uma melhor continuidade do trabalho de pesquisa e maior rapidez em termos de aplicação pelos grupos de pessoas interessadas.

Criação de um grupo de pesquisa para que possibilite uma maior interação entre outros pesquisadores, que outros vários projetos necessários para preenchimento de várias lacunas existentes possam ser feitos mais rapidamente e de forma simultânea e que possibilite a orientação e formação de outros pesquisadores para possa garantir a continuidade e o crescimento em termos de trabalhos essa linha de pesquisa.

Planeja-se que os resultados da investigação realizada durante este projeto sejam publicados em periódicos nacionais e internacionais e em anais de eventos científicos, e que também sejam apresentados em conferências.

Fazer uma avaliação crítica construtiva do questionário completo elaborado para o Censo 2010 pelo IBGE com a finalidade de aperfeiçoar a pesquisa de investigação das pessoas com deficiência para o próximo censo.

## **5 – Palavras chaves**

Cruzamentos entre variáveis, Teste de homogeneidade, Censo demográfico 2010, Análise exploratória, Seleção de variáveis, Seleção de modelos, Geoestatística, Risco Deficiência, Modelagem de equações estruturais, Cadastro Deficiência, Grupo de Pesquisa.

## **6 – Justificativas para o projeto**

Poucas pesquisas feitas sobre pessoas com deficiência, e, menos ainda que levem em conta seus aspectos quantitativos e qualitativos em que relacionam os tipos de deficiência com outras variáveis como as que são relacionadas com moradia e infraestrutura de entorno.

Necessidade de traçar relacionamento entre as diferentes deficiências e outras variáveis explicativas ligadas a saúde, educação, trabalho, moradia, lazer, saúde, posse de determinados bens, infraestrutura urbana e outros. Isto tem por finalidade de oferecer melhor suporte para essas pessoas ligadas a um seguinte tão complexo e muitas vezes carentes em termos de acessibilidade.

Ausência de estudos comparativos entre pessoas com deficiência e que não apresentam deficiência.

Grande desconhecimento de como vivem, onde moram e quantas são as pessoas com deficiência e o seu relacionamento com o entorno.

Necessidade de existência de atualização em tempo real para que possa ser acompanhado pelas pessoas interessadas, as condições que diz respeito da qualidade de vida das pessoas com deficiência.

Implementação do risco deficiência que possa ser calculado, monitorado e a sua evolução acompanhada de forma contínua e real.

Avaliar a situação e distribuição das residências das pessoas com deficiência ao longo do país, bem como, avaliar a infraestrutura do entorno em termos de risco e de benefícios para as pessoas com deficiência em termos de acessibilidade, de serviços a serem oferecidos, e



também, problemas decorrentes devido a inacessibilidade ou difícil acesso e inexistência de certas componentes que poderiam melhorar suas condições de vida.

Este projeto de pesquisa propõem o ajuste das diferentes deficiências em relação a variáveis ligadas a identificação, instrução, família, posse de outros bens e condições de moradia utilizando para ajuste MEE; implementação do risco deficiência utilizando escores fatoriais do MEE, regressão logística binária, regressão logística ordinal e regressão logística estereótipo; modelagem geoestatística para que possa melhor avaliar as condições de entorno e infraestrutura das moradias das pessoas com deficiência; implementação de cadastro deficiência disponibilizado na internet em tempo real e contínuo sob condição de cadastro gratuito por parte dos interessados; e, finalmente; formação de um grupo de pesquisa denominado a priori **Grupo de Estudos Quantitativos, Qualitativos e Comparativos da Discapacidade** para que melhor possibilidade a participação de outros pesquisadores, desenvolvimento de outros projetos, orientação e formação de novos pesquisadores, participação e publicação de artigos em revistas e eventos científicos ligados a essa área a nível nacional e internacional.

## **7 - Materiais e métodos**

Para este projeto de pesquisa, planeja-se utilizar a infraestrutura existente no Instituto de Matemática e Estatística da Universidade de São Paulo, bem como equipamentos adicionais a serem adquiridos com recursos obtidos para o presente projeto.

Pretende-se utilizar micro computador com sistema operacional windows, editor de texto, planilha eletrônica do tipo excel, banco de dados, programa de apresentação do tipo power point ou similar.

A princípio, pretendemos utilizar programas estatísticos que suportam a leitura de grandes bancos de dados como SAS, Epi-info, R e SPSS, mas com o tempo, pretendemos migrar tudo para o R ([www.r-project.org](http://www.r-project.org)) que é de domínio público, diminuindo assim as despesas com aquisição de pacotes estatísticos comercializados. Neste sentido, pretendemos implementar programas próprios na linguagem R para a continuidade da execução deste projeto.

Outro ponto importante é que o banco de dados que pretendemos utilizar nesta pesquisa encontra-se disponível no site do IBGE ([www.ibge.gov.br](http://www.ibge.gov.br), dados do censo 2010 selecionados para compor a amostra).

Para a criação do cadastro e de formulário para coleta dos dados sobre pessoas com deficiência, pretendemos utilizar aplicativos como Epi-info localizado em

<http://www.cdc.gov/epiinfo/7/index.htm> e, que pode ser baixado de forma gratuita, em seu módulo de criação de formulário.

## **8 - Descrição do estudo**

Trata-se de um estudo observacional em que a coleta de dados foi realizada em 2010 pelos recenseadores do IBGE. A unidade amostral considerada é indivíduo no caso de pessoas e moradia no caso de domicílios. Foram pesquisados 20800804 pessoas distribuídas em 6191924 domicílios sorteados da população para responder ao questionário completo que era aplicado e preenchido pelo recenseador.

### **8.1- Descrição das variáveis**

Neste projeto serão considerados as seguintes variáveis alocadas em tópicos a seguir: identificação: número de identificação, município, estado, região, sexo, situação de domicílio, idade, forma de declaração da idade, raça, registro de nascimento, naturalidade e nacionalidade, nasceu neste município, nasceu nesta unidade da federação, UF ou país estrangeiro de nascimento, tempo de moradia na UF, tempo de moradia no município; deficiência: número de deficiências, dificuldades permanentes para enxergar, ouvir, locomover e mental; instrução: saber ler e escrever, frequenta escola ou creche, curso que frequenta atualmente, série ou ano que frequenta, série que frequenta, outra graduação, curso mais alto que frequentou e nível de instrução; família: relação de parentesco ou de convivência com a pessoa responsável pelo domicílio, ordem lógica, vive em companhia do cônjuge ou companheiro(a), número de ordem do cônjuge ou companheiro, situação de parentesco do respondente, natureza da união, estado civil e número de filhos; condições de moradia: situação de ocupação, condição de ocupação, quantas pessoas moravam neste domicílio, ocorrência de óbito, responsável pelo domicílio, tipo de moradia, valor de aluguel, tipo de unidade doméstica, material permanente de paredes externas, número de cômodos, densidade morador/cômodo, número de dormitórios, densidade morador/dormitório, número de banheiros, sanitários, esgotamento sanitário, forma de abastecimento de água, canalização, destino do lixo, existência de medidor ou relógio de energia elétrica e energia elétrica; outros bens existentes na residência: rádio, televisão, máquina de lavar, geladeira, telefone celular, telefone fixo, microcomputador, acesso à internet, moto e carro, e por fim; trabalho: quantos trabalhos tinha, nesse trabalho era, quantas pessoas empregava, contribuía para a previdência oficial, rendimento bruto ou retirada mensal, rendimento no trabalho principal, rendimento nos outros trabalhos, rendimento em todos os trabalhos, retorna do trabalho para casa diariamente, tempo de deslocamento entre casa e trabalho, quantas horas trabalhava por semana no emprego principal, aposentadoria, bolsa família, outros programas sociais ou

transferências, outras fontes, rendimento domiciliar em salários mínimos, rendimento domiciliar em salários mínimos per capita, situação de ocupação, posição de ocupação no emprego principal e posição no emprego secundário.

Para este estudo estamos propondo, também, a criação das seguintes variáveis com as suas respectivas categorizações: no tópico trabalho: rendimento em salários mínimos (sm, que na época da realização do Censo Demográfico 2010 era de 510 reais) categorizados como 1, se ganhar de 0 até 1 sm; 2, de 1 até 3 sm; 3, de 3 até 7 sm; 4, de 7 até 15 sm, e por fim; 5, de 15 sm ou mais; no tópico identificação: identificação (número do questionário ou indivíduo) e idade categorizada (1, se tiver de 0 até 15 anos; 2, se tiver entre 15 até 65 anos, e por fim; 3, de 65 anos ou mais, no tópico família: número de filhos categorizados (1 para sem filho; 2, para número de filhos entre 1 e 2; 3, para número de filhos entre 3 a 5, e por fim; 4, número de filhos a partir de 6 ou mais), e por fim; no tópico deficiência: deficiências (quantidade de deficiências que cada indivíduo possui, e varia de 0 até 4), defic1 (0 se não possui deficiência e 1 se possui pelo menos uma deficiência), aluguel em salários mínimos categorizados (1, para valores de aluguéis entre 0 e 0,5sm; 2, para valores entre 0,5 e 1sm; 3, para valores entre 1 e 2sm; 4, para valores entre 2 e 4sm, e, por fim; 5, para valores maiores que 4sm), número de cômodos categorizados (1, se o número for entre 1 a 3 cômodos; 2, se for 4 cômodos; 3, se for 5 cômodos; 4 se for 6 cômodos; 5, se for 7 cômodos, e, por fim; 6, se a casa tiver 8 cômodos ou mais), número de dormitórios categorizados (1, se a residência possui um dormitório; 2, se possui 2 dormitórios; 3, se a residência possui 3 dormitórios; 4, se a residência possui 4 dormitórios ou mais), densidade morador por cômodo categorizado (1, se a densidade estiver entre 0 e 1 morador por cômodo; 2, se estiver entre 1 e 2 moradores por cômodo, e, por fim; 3 se a residência possuir uma densidade acima de 2 moradores por cômodo) e densidade morador por dormitório categorizado (1, se a residência tiver entre 0 e 1 moradores por dormitório; 2, se tiver entre 1 e 2; 3, se tiver entre 2 e 3; 4, se tiver entre 3 e 4; 5, se tiver entre 4 e 5, e, por fim; 6 se a residência tiver 5 moradores por dormitório ou mais).

## **9 – Análise exploratória**

A busca de respostas a perguntas nas áreas de saúde e deficiência demanda muitas vezes a coleta de grande quantidade de dados para análise (Medronho et al., 2009). Extrair informações de grandes volumes de dados demanda técnicas específicas tanto para a coleta, organização e síntese destes dados, quanto para a análise propriamente dita. Esta fase corresponde a etapa do processo da análise conhecida por análise exploratória de dados, que trata-se de um conjunto de técnicas que objetivam descrever, analisar e interpretar dados numéricos e amostra. Para este projeto pretendemos descrever cada variável separadamente

(descrição na forma univariada, exemplo, variável enxergar e idade); relacionamento entre duas variáveis (descrição na forma bivariada, exemplo, relacionamento entre enxergar e nível de instrução, e, enxergar e idade), e, por fim; relacionamento entre três variáveis ou mais (exemplo, estudo para cada sexo do relacionamento entre, enxergar e nível de instrução).

### 9.1 – Seleção de dados

As variáveis selecionadas podem ser do tipo categórica ou do tipo quantitativa. As variáveis categóricas assumem valores finitos, diferem na forma e podem ser nominais ou ordinais. Ao contrário da variável nominal, existe uma ordem entre os possíveis valores de uma variável categórica ordinal. Exemplos de variáveis nominais são sexo (masculino, feminino), estado civil (solteiro, casado, divorciado, desconhecido). Exemplos de variáveis ordinais são nível de instrução (sem instrução e fundamental incompleto, fundamental completo e segundo grau incompleto, segundo grau completo e superior incompleto e superior completo ou mais). E enxergar (sim, não consigo de modo algum; sim, mas com muita dificuldade, sim, mas com um pouco de dificuldade, e, por fim; não apresento nenhuma dificuldade).

As variáveis quantitativas assumem valores numéricos e podem ser do tipo contínua (os possíveis valores são numéricos números reais), exemplo, idade ou discretas (os possíveis valores fazem parte de um conjunto finito), exemplo, número de filhos. As variáveis selecionadas são as que serão utilizadas para o desenvolvimento das pesquisa nos mais diversos procedimentos.

Neste trabalho, nota-se no banco de dados a ser utilizado na análise, um predomínio de variáveis categóricas ordinais seguido de categóricas não ordinais.

### 9.2 - Teste de homogeneidade

Esse teste consiste em verificar se uma variável aleatória se comporta de modo homogêneo, em várias subpopulações, e, fixa o tamanho da amostra em cada uma destas subpopulações e, então, seleciona uma amostra de cada uma delas.

Para o cálculo dos valores esperados, supondo a existência de homogeneidade entre as subpopulações, utiliza-se para cada casela  $(i, j)$ , conforme expressão a seguir:

$$e_{i,j} = n_i \times \frac{\text{total da coluna } j}{\text{total geral}} \quad (1)$$

O total da linha  $n_i$  indica o tamanho da amostra da subpopulação  $i$ , ao passo que o quociente, total da coluna  $j$  dividido pelo total geral, representa a proporção de ocorrências do

valor da variável correspondente à coluna  $j$ . Caso haja homogeneidade do comportamento da variável correspondente, espera-se que essa proporção seja a mesma, em todas as subpopulações. O próximo passo é calcular  $Q^2$  que é a diferença entre os valores observados e esperados utilizando a expressão abaixo:

$$Q^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{o_{ij} - e_{ij}}{e_{ij}} \quad (2)$$

onde  $r$  e  $s$  representam o número de linhas e de coluna respectivamente.

Para interpretar a expressão de  $Q^2$ , nota-se que o termo  $o_{ij} - e_{ij}$  indica a diferença entre o valor observado e o valor esperado na linha  $i$  e coluna  $j$ , se houvesse homogeneidade (Magalhães and Lima, 2011).

Para um número grande de observações, a expressão de  $Q^2$  é qui-quadrado com  $(r - 1)(s - 1)$  graus de liberdade. A região crítica contém grandes valores de  $Q^2$ , isto é,  $RC = Q^2: Q^2 \geq q_c$ , com  $q_c$  sendo determinado pelo nível de significância do teste, ou seja,  $\alpha = P(Q^2 \geq q_c / H_0 \text{ é verdadeiro})$ .

Para este estudo, utilizaremos testes de homogeneidade com a finalidade de verificar se uma variável aleatória se comporta de modo similar, ou homogêneo, em várias subpopulações. Exemplo, suponhamos se estamos interessados em saber se o nível de instrução se altera com os níveis de deficiência visual. Consideremos 20800804 pessoas, selecionadas para compor a amostra, entrevistadas pelos recenseadores do IBGE para o Censo Demográfico 2010.

## 10 – Amostragem

Com a dificuldade de fazer as análises desejadas para um conjunto de mais de 20 milhões de casos, optamos trabalhar com amostragem probabilística e complexa, ou seja, extrair um subconjunto desses dados que seja o melhor representativo possível do conjunto original, ou seja, vamos trabalhar com uma sub amostra e fazer as análises a partir desta sub amostra que mantenha as mesmas proporções por estado, região, sexo, raça e município que será utilizada para aplicação da modelagem de equações estruturais.

Estudos por amostragem são aqueles que obtêm uma ou mais amostras de uma população de interesse, com a finalidade de estimar parâmetros populacionais. É importante que os diferentes procedimentos amostrais satisfaçam aos seguintes critérios (Mundstock, 2005): que as amostras sejam representativas da população; que forneçam estimativas

precisas das características da população, podendo medir sua confiabilidade; e por fim, que tenham pequeno custo de seleção.

A amostragem pode ser classificada em probabilística e não probabilística, sendo que na probabilística, é possível calcular com antecedência, a probabilidade de obter-se cada uma das amostras possíveis, sendo que todas as unidades da população tem probabilidade maior que zero de participar da amostra. É importante observar que a aleatoriedade da amostra depende do processo pelo qual ela é obtida. Nas amostras probabilísticas é possível estimar, com uma determinada probabilidade, os erros de amostragem ou as discrepâncias entre as estimativas amostrais e os valores populacionais que seriam obtidos observando todas as unidades da população (Cochran, 1977).

Ao contrário, a amostragem não probabilística é um procedimento pelo qual não se pode associar probabilidade de seleção de unidades e, conseqüentemente, não é possível determinar a confiabilidade dos resultados das amostras em termos probabilísticos. Consideraremos neste projeto apenas procedimentos de estimação adequados à amostras selecionadas de forma probabilística.

Para que um plano amostral seja considerado complexo deve apresentar características como: estratificação com a população subdivida em estratos (subpopulações mutuamente exclusivas) e consiste em selecionar uma amostra em cada estrato e combinar estas amostras numa única amostra para estimar parâmetros da população, tendo como vantagem o aumento da precisão das estimativas, possibilidade de obtenção de informação em nível de estrato e facilidade de coleta de dados, por razões físicas ou administrativas (Silva, 1998); conglomeração, a população é dividida em  $M$  grupos ou conglomerados que servem como unidades primárias de amostragem (UPA) de maneira que cada unidade da população é associada a um e somente um conglomerado. Cada conglomerado é formado por  $N_i$  unidades secundárias de amostragem. Dos  $M$  conglomerados na população é selecionada uma amostra de tamanho  $m$ .

Segundo Bussab e Bolfarine (2005), a amostragem por conglomerado é muito usada em populações humanas, onde frequentemente são sorteadas cidades, depois os bairros, os domicílios e, finalmente, os moradores. Uma das desvantagens para o uso da amostragem de conglomerados prende-se ao fato de que as unidades, dentro de um mesmo conglomerado, tendem a ter valores parecidos em relação às variáveis que estão sendo pesquisadas, e isso torna este tipo de amostragem menos eficiente. Por fim, a amostragem complexa também pode ser caracterizada pela amostragem com probabilidades variáveis que em alguns

procedimentos amostrais, algumas unidades da população são “mais importantes” por terem uma contribuição maior no valor do parâmetro, neste caso estabelecem-se probabilidades desiguais de seleção às diferentes unidades da população (Natarajan et al., 2008).

Nos casos em que a probabilidade de seleção é proporcional à uma medida de tamanho da população, o procedimento amostral é definido como amostragem com probabilidade proporcional ao tamanho. A vantagem em selecionar a unidade em probabilidade proporcional ao tamanho é ter uma amostra mais representativa da população, e assim, aumentar a precisão dos estimadores (Mundstock, 2005).

Para este projeto, pretendemos identificar variáveis importantes de estratificação, calcular tamanho mínimo da amostra necessária para que possa obter representatividade, e, por fim; validar observando as estimativas.

## **11 - Imputação**

Segundo Barroso (1995) a não resposta pode ser total ou parcial: total, quando todas as informações são perdidas, e parcial, quando são perdidas uma parte delas.

Segundo Rubin (1976) e Laird (1988), se o mecanismo de não resposta não depende da variável de interesse, dizemos que os dados são perdidos por um mecanismo aleatório MCAR (do inglês, *Missing Completely at Random*). Quando a probabilidade de não resposta depende dos dados presentes mas não dos ausentes, os dados são ditos perdidos por um mecanismo aleatório MAR (do inglês, *Missing at Random*) e o mecanismo de não resposta é ignorável.

O processo de não resposta pode ainda ser não ignorável quando a falta dos dados não é aleatória NMAR (do inglês, *Not Missing at Random*). Esse é o caso em que a probabilidade de não resposta depende dos dados ausentes, no conjunto de dados que pretendemos utilizar, o melhor exemplo é renda, pois, na prática tem mostrado que a perda de informações é maior entre as pessoas de maior renda.

Neste estudo, com exceção da variável renda, cuja perda de informação deve ser tratada como não ignorável, nos outros casos serão tratados como ignorável, por não ser necessário especificar um modelo resposta.

Segundo Poletto (2011), métodos estatísticos específicos desenvolvidos para lidar com o problema de dados com omissão podem ser grosseiramente agrupados em três classes: i) extensão dos métodos tradicionais de estimação em pesquisas de levantamento de dados com utilização de pesos de tal forma que procure manter a proporcionalidade entre os níveis de

cada variável para as variáveis categóricas; ii) substituir os dados faltantes por estimativas com o intuito de utilizar os métodos tradicionais de análises para a análise de dados completos que é a opção a ser considerada neste trabalho, e, por fim; iii) considerar um mecanismo gerador de dados omissos.

Para este projeto, os valores dos respondentes, similares em relação às variáveis auxiliares, serão selecionados para a imputação, doando características definidoras para os dados. Assim, localizaremos o indivíduo com o dado observado mais parecido com o indivíduo com dado faltante em relação às variáveis auxiliares e substitui-se tal dado faltante pelo valor do respondente pareado. Havendo mais de um respondente pareado, deverá ser usado o método de imputação do “vizinho mais próximo”, onde algum critério de classificação é desenvolvido para determinar o registro mais semelhante àquele com o dado faltante e determinado registro se torna “doador” desses dados. Um exemplo de variáveis, doadoras para determinar o perfil usado na determinação de tal imputação, são as variáveis dos blocos identificação e deficiências, como sexo, idade, região, estado, município, situação de domicílio, raça, nacionalidade, naturalidade, deficiência para enxergar, ouvir, locomover e mental. Assim, verifica-se qual indivíduo respondente tem o mesmo perfil do não-respondente em relação a tais dados. Aquele do mesmo padrão, ou perfil, é o doador. Desta forma, o dado faltante no item será preenchido com a resposta do doador.

## **12 – Equações Estruturais**

A MEE é uma metodologia que possibilita analisar várias relações de dependência entre si e aceita a incorporação de variáveis que não podem ser mensuradas diretamente,

O MEE é uma ferramenta analítica relativamente nova, mas suas raízes se remontam à primeira metade do século XX, o seu desenvolvimento resultou da evolução do modelo de multi-equações desenvolvidas principalmente na econometria e também dos princípios de mensuração da psicologia e sociologia (Black, 1962; Habelmo, 1943; Wright, 1921). A ideia de combinar as características econométricas e psicométricas em, um único modelo matemático surgiu no verão de 1970 e foi inspirado no trabalho de Golberger (1971). O termo análise de estrutura de covariância foi introduzido para descrever o que atualmente é chamado de AFC (Latif, 2000; Bock and Bargmann, 1989).

A MEE possui duas partes: i) modelo de mensuração formada pelas equações (3) e (4) e ii) modelo estrutural formado pela equação (5). Enquanto o modelo de mensuração especifica como as variáveis latentes são mensuradas pelas variáveis observadas, descrevendo sua validade e confiabilidade, o modelo estrutural especifica as relações de causa e efeito



entre as variáveis latentes, apresentando os efeitos causais e o total da variância não explicada (Bistaffa, 2010).

Em termos matemático a MEE pode ser escrita da seguinte forma:

i) Modelo de mensuração:

$$y_{(p \times 1)} = A_y \eta_{(m \times 1)} + \varepsilon \quad (3)$$

$$x_{(q \times 1)} = A_x \xi_{(n \times 1)} + \delta \quad (4)$$

em que (3) é a equação de mensuração das variáveis latentes endógenas, e (4), a das variáveis latentes exógenas;  $A_{y_{(p \times m)}}$  e  $A_{x_{(q \times n)}}$  são matrizes de parâmetros, que representam os efeitos das variáveis latentes (respectivamente endógenas e exógenas) sobre seus respectivos indicadores, e, por fim;  $\varepsilon_{(p \times 1)}$  e  $\delta_{(q \times 1)}$  são, nesta ordem, os vetores de erros de mensuração para  $\mathbf{y}$  e  $\mathbf{x}$ . Para que possamos interpretar completamente estes parâmetros é necessário associar uma escala à variável latente. Geralmente associa-se à escala da variável latente a um dos indicadores ou a variância da variável latente.

ii) Equações estruturais:

Sendo os vetores  $\eta_{(m \times 1)}$  e  $\xi_{(n \times 1)}$  de variáveis latentes endógenas e exógenas respectivamente, então a equação estrutural do MEE pode ser escrita como:

$$\eta_{(m \times 1)} = B_{(m \times m)} \eta_{(m \times 1)} + \Gamma_{(m \times n)} \xi_{(n \times 1)} + \zeta_{(m \times 1)}, \quad (5)$$

em que  $B_{(m \times m)}$  é uma matriz de parâmetros que representam os efeitos das variáveis latentes endógenas sobre os outros construtos;  $\Gamma_{(m \times n)}$  é uma matriz de parâmetros que representam o efeito das variáveis exógenas sobre os outros construtos, e, por fim;  $\zeta_{(m \times 1)}$  vetor aleatório de erros.

Para MEE são consideradas as seguintes suposições:  $\xi$  e  $\zeta$  são não correlacionados; I-B é não singular;  $E[\zeta] = 0$ ;  $E[\eta] = 0$ ;  $E[\xi] = 0$ ;  $\varepsilon$  é não correlacionado com  $\eta$ ;  $\delta$  é não correlacionado com  $\xi$ ;  $\varepsilon$  e  $\delta$  são não correlacionados entre si, mas podem ter suas respectivas variáveis correlacionadas entre si, e, por fim;  $\zeta$ ,  $\varepsilon$  e  $\delta$  são mutuamente não correlacionadas, com as seguintes matrizes de covariância:  $\Phi_{(n \times n)} = E[\xi \xi']$ ;  $\Psi_{(m \times m)} = E[\zeta \zeta']$ ;  $\Theta_{\varepsilon_{(p \times p)}} = E[\varepsilon \varepsilon']$ , e, por fim;  $\Theta_{\delta_{(q \times q)}} = E[\delta \delta']$ .

As suposições que uma variável latente endógena não tem influência direta sobre ela mesma, o que leva a diagonal principal da matriz  $\mathbf{B}$  ser nula.

É possível mostrar que a matriz de covariância de  $Z_{(p+q) \times 1} = (y^T, x^T)^T$  de dimensão  $(p + q) \times (p + q)$  para o modelo completo é dado por:

$$\Sigma(\theta) = \begin{pmatrix} A_y(I-B)^{-1}(\Gamma\Phi\Gamma^T + \Psi)(I-B^T)^{-1}A_y^T + \Theta_\varepsilon & A_y(I-B)^{-1}\Gamma\Phi A_x^T \\ A_x\Phi\Gamma^T(I-B^T)^{-1}A_y^T & A_x\Phi A_x^T + \Theta_\delta \end{pmatrix} \quad (6)$$

em que,  $\theta$  é um vetor de dimensão  $t \times 1$  de parâmetros livres (coeficientes que acompanham as variáveis latentes nas equações estruturais e de mensuração, variâncias e covariâncias das variáveis latentes independentes, erros de mensuração e erros das equações estruturais);  $\Phi_{(n \times n)}$  é a matriz de covariâncias das variáveis latentes exógenas (ou seja, das  $\zeta$ 's);  $\Psi_{(m \times m)}$  é a matriz de covariâncias dos erros nas equações (isto é, dos  $\zeta$ 's);  $\Theta_{\varepsilon_{(p \times p)}}$  é a matriz de covariância de  $\varepsilon$ ;  $\Theta_{\delta_{(q \times q)}}$  é a matriz de covariância de  $\delta$ , por fim; os demais casos são conforme já definidos anteriormente.

As suposições que foram violadas, por incorporarem variáveis categóricas são  $E[\xi] = 0$ ;  $\varepsilon$  é não correlacionado com  $\eta$ ; e, por fim;  $\eta$  é da forma ajustada. As consequências são: a inexistência de igualdade em (3) fazendo com que o modelo de mensuração não seja válido para  $y$ ; a distribuição das variáveis ordinais geralmente difere da distribuição da variável latente contínua;  $aCov(s_{ij}, s_{gh})$  difere de  $aCov(s_{ij}^*, s_{gh}^*)$ , e, por fim; violação da estrutura de covariância tirando a igualdade da expressão (15).

Uma vantagem da MEE é a sua interface gráfica, isto é, a possibilidade de representações gráficas das relações existentes entre as variáveis observadas e as latentes e das correlações entre os parâmetros através de diagramas de caminho, cuja construção pode ser realizada da seguinte maneira: variáveis observadas são representadas por quadrados ou retângulos; as variáveis latentes, por círculos ou elipses; as variáveis de erro podem ou não ser representadas dentro das figuras; flechas retas indicam influência direta; flechas retas duplas, influência recíproca, e por fim; flechas curvas, possíveis correlações entre as variáveis (Latif, 2000).

Outra vantagem é que MEE possibilita a estimação simultânea de uma série de equações múltiplas distintas, mas que se inter-relacionam (Vieira et al., 2008).

A hipótese fundamental de um MEE é que a matriz de covariância das variáveis observadas é função dos parâmetros do modelo. Se o modelo for correto e se conhecem os verdadeiros valores dos parâmetros, a matriz de covariância populacional é reproduzida exatamente.

Conforme mencionado, uma das dificuldades deste estudo é que o conjunto das variáveis que serão utilizadas nesta análise é constituída por variáveis categóricas, sejam ordinais como as variáveis que caracterizam as diferentes deficiências ou não ordinais como por exemplo raça. Em vista disso, serão feitos procedimentos corretivos como determinar os limiares discretizantes (Bistaffa, 2010) e em seguida calcular a correlação entre os diferentes pares de variáveis utilizando os métodos de correlação especial como policórica (para os casos em que as duas variáveis são ordinais com três categorias ou mais), biserial (uma variável é contínua e a outra é categórica ordinal com apenas duas categorias), tetracórica (as variáveis são ordinais e dicotômicas), poliserial (uma variável é contínua e a outra é ordinal com três categorias ou mais), phi (duas variáveis nominais dicotômicas) e lambda (entre variáveis nominais e ordinais e entre variáveis nominais com pelo menos uma delas com três categorias ou mais) compondo uma matriz de correlação híbrida em que a correlação para cada par de variáveis dependerá do tipo que está sendo considerada.

Neste trabalho temos como variáveis contínuas: idade, número de filhos e renda; variáveis nominais com duas categorias: sexo e situação do domicílio; nominais com três categorias ou mais: estado, região, raça, registro do nascimento, naturalidade, nacionalidade, natureza da união, estado civil, posição no emprego principal, posição no emprego secundário e frequenta escola ou creche; ordinais com duas categorias: saber ler e escrever, outra graduação, se retorna do trabalho para casa diariamente, condição de ocupação e situação de ocupação, e, por fim; ordinal com mais que duas categorias: dificuldades permanentes para enxergar, ouvir, locomover e mental, nível de instrução, curso mais alto que frequentou, e tempo de deslocamento entre casa e trabalho.

### 12.1 – Correlações especiais

Os tipos de correlação entre duas variáveis que serão utilizadas são as seguintes:

- **policórica** para os casos que temos duas variáveis ordinais com 3 categorias ou mais.

Esta correlação é resolvida por métodos numéricos igualando o quadrado da correlação de Pearson com a série policórica truncada. A estimativa da correlação policórica é dada por

$$\phi^2 = \left[ \chi^2 - (r-1)(s-1) \right] / n = \sum_{i=1}^{r-1} \sum_{j=1}^{s-1} \left[ \sum_{m=1}^{\infty} a_m b_{jm} \rho^m \right]^2, \quad (7)$$

onde  $\chi^2$  é o valor do qui quadrado para a tabela de contingência  $r \times s$ .

- **tetracórica** nos casos em que as duas variáveis são dicotômicas e ordinais. Uma possível expressão para correlação tetracórica é dada por:

$$\hat{\rho}_t = \cos\left(\frac{180\sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}\right), \quad (8)$$

em que  $a, b, c$  e  $d$  são as frequências da tabela  $2 \times 2$ , e, por fim;  $\hat{\rho}_t$  é o estimador da correlação tetracórica.

- **poliserial** nos casos que temos uma variável contínua e outra ordinal com três categorias ou mais. Uma possível expressão para seu cálculo é dada por

$$\hat{\rho}_{ps} = \frac{\rho}{\sigma_D} \sum_{j=1}^{s-1} \phi(\tau_j)(d_{j+1} - d_j), \quad (9)$$

em que  $\hat{\rho}_{ps}$  é o estimador da correlação poliserial;  $\sigma_D$  é a variância de  $D$ ; e,  $\phi(\tau_j) = (2\pi)^{-1/2} \exp\{-\tau_j^2/2\}$  é a densidade da distribuição normal padrão no ponto  $\tau_j$ .

- **bisserial** para os casos em que uma variável é contínua e a outra é ordinal com duas categorias. Uma possível expressão para este caso é dada por:

$$\hat{\rho}_{bs} = \frac{pq(\bar{x}_1 - \bar{x}_0)}{\sigma_X \phi(\tau)}, \quad (10)$$

em que  $\hat{\rho}_{bs}$  é o estimador da correlação biserial;  $p$  é a proporção de valores iguais a 1;  $q = 1 - p$ ;  $\bar{x}_1$  é a média dos  $x$ 's da amostra  $D = 1$ ;  $\bar{x}_0$  é a média dos  $x$ 's da amostra  $D = 0$ ;  $\sigma_X^2$  é a variância de  $X$ ;  $\tau = \phi^{-1}(q)$ , e, por fim;  $\phi(\tau_j) = (2\pi)^{-1/2} \exp\{-\tau_j^2/2\}$ .

- **eta** para uma variável quantitativa e outra categórica nominal. Uma expressão utilizada para cálculo é dada por

$$\hat{\eta} = \sqrt{\frac{\text{soma de quadrados entre grupos}}{\text{soma de quadrados total}}}. \quad (11)$$

em que  $\hat{\eta}$  é o estimador da correlação eta.

- **contingência** para duas variáveis nominais. Uma expressão para o cálculo é dada por

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}, \quad (12)$$

em que  $C$  é o coeficiente de contingência;  $\chi^2$  é o qui-quadrado calculado para os dados, e, por fim;  $n$  é o número de elementos da amostra.

- **phi** para as duas variáveis dicotômicas podendo ou não ser nominais. Uma possível expressão que pode ser utilizada em seu cálculo é dada por

$$\hat{\phi} = \frac{(ad - bc)}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}. \quad (13)$$

em que  $\hat{\phi}$  é o estimador da correlação phi e as demais são conforme já mencionadas anteriormente.

- **Pearson** para duas variáveis contínuas. Uma possível expressão é dada por:

$$\hat{\rho} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (14)$$

em que  $\hat{\rho}$  é o estimador da correlação de Pearson, X e Y são variáveis contínuas.

Nota-se que para esta pesquisa comporão a matriz de correlação híbrida os valores das correlações obtidas pelos diferentes métodos de acordo com o tipo de cada variável que compõem cada um dos pares das diferentes combinações entre as vinte e nove variáveis duas a duas até que a matriz de correlação seja completada.

Um possível modelo utilizando MEE considerando os dados do IBGE é apresentado na Figura 4.

Nota-se que o modelo proposto na Figura 4 é formado por variáveis observadas (as variáveis que pertencem aos diferentes grupos identificação, instrução, família, condições de moradia, outros bens existentes na residência, trabalho e deficiência), latentes (variáveis que estão agrupadas em risco) e deficiências, nota-se que as variáveis observadas podem ser relacionadas a risco de cada indivíduo que foi entrevistado de ser ou não pessoa com deficiência.

## 12.2 - Teste de hipóteses

A hipótese fundamental de um MEE é que a matriz de covariância das variáveis observadas é função dos parâmetros do modelo. Se o modelo for correto e se conhecem os verdadeiros valores dos parâmetros, a matriz de covariância populacional é reproduzida exatamente. A equação (15) mostra a formalização desta hipótese (Vieira and Skinner, 2008; Vieira, 2009):

$$\Sigma = \Sigma(\theta), \quad (15)$$

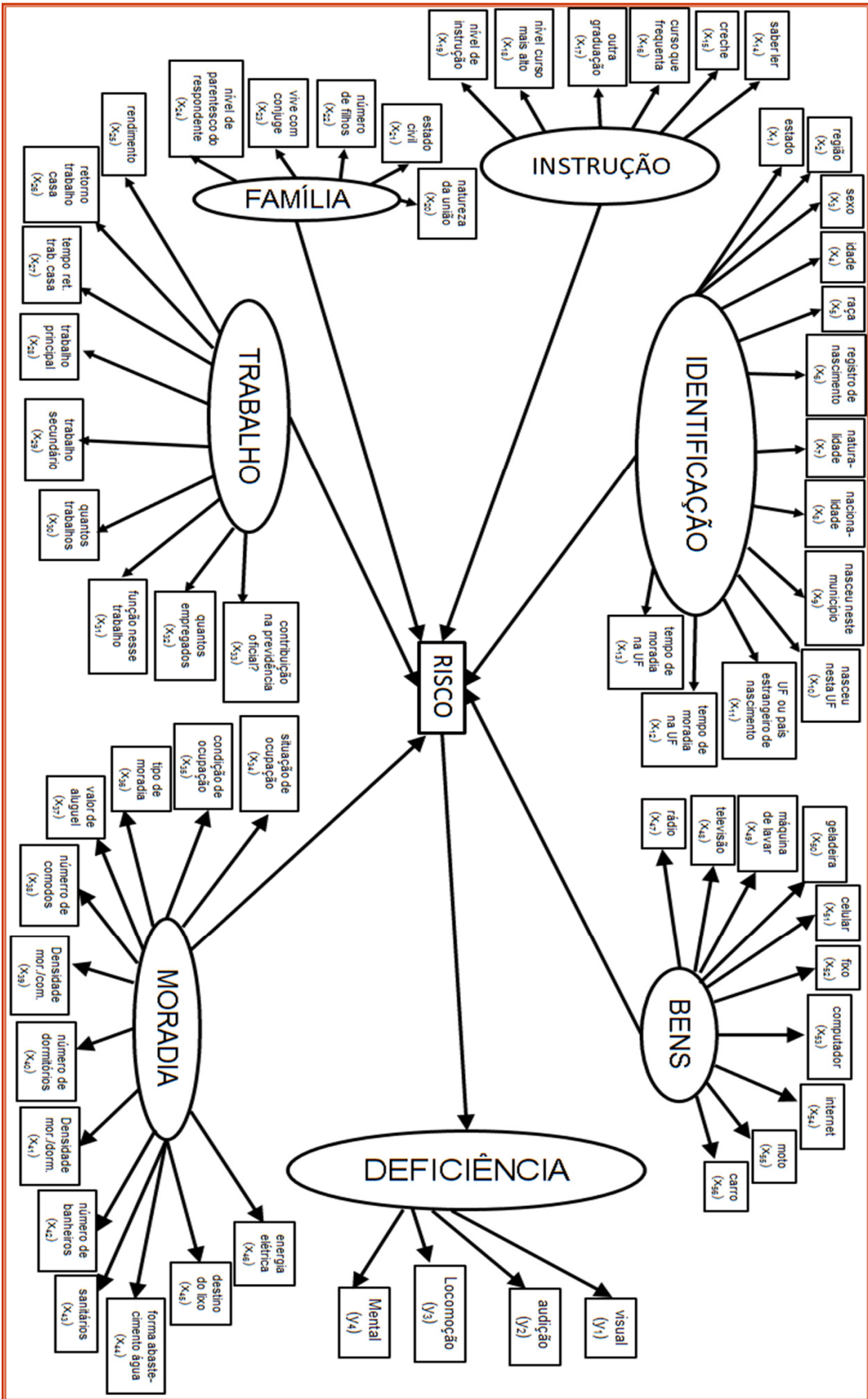


Figura 4. Modelo proposto para o projeto

em que  $\Sigma$  é a matriz de covariância populacional das variáveis observadas ( $X$ 's),  $\theta$  é o vetor de parâmetros do modelo, e  $\Sigma(\theta)$  implicada pelo modelo. De tal modo que, para o cálculo das medidas de ajuste para o modelo, foi considerado os estimadores  $S$  e  $\Sigma(\hat{\theta})$ , onde  $\Sigma(\hat{\theta})$  é a matriz de covariância avaliada em  $\hat{\theta}$  (ou matriz de covariância, como sugerida por (Jöreskog and Sörbom, 1989; Jöreskog and Sörbom, 1997)).

A maioria das medidas de ajuste de modelo proposto na literatura consideram funções de  $S$  e  $\Sigma$ . Bollen (1989) discute algumas limitações das medidas de ajuste global como por exemplo: i) estas medidas não são aplicáveis para modelos identificados exatamente (Como  $S$  e  $\Sigma(\hat{\theta})$ , nesta circunstância); e, ii) esta deve contradizer o ajuste de partes do modelo (por exemplo, o ajuste global deve ser favorável, mas, estimativas dos parâmetros podem não ser estatisticamente significantes). Em comparação entre os diferentes desempenhos, a qualidade do teste de ajuste deve ser inicialmente uma hipótese nula e uma hipótese alternativa respectivamente (Browne, 1982; Satorra and Bentler, 1986; Satorra and Bentler, 1988; Satorra, 1989; Satorra and Bentler, 1994).

A solução de máxima verossimilhança (MV) também nos dá a vantagem de um teste estatístico de qualidade de ajuste do modelo. Podemos comparar o ajuste de um modelo particular (modelo restrito) com um modelo mais geral (modelo composto, cuja as restrições são menos rigorosas que no modelo restrito).

Podemos testar a adequação de qualquer modelo MEE usando a Razão de Verossimilhança. Essa razão toma a seguinte forma:

$$-2[\ln(L_R) - \ln(L_F)] \sim \chi^2, \quad (16)$$

onde  $L_R$  é a verossimilhança do modelo completo e  $L_F$  é a verossimilhança do modelo restrito. O número de graus de liberdade para o teste é a diferença entre os números de graus de liberdade para os respectivos modelos e quanto maior o valor da estatística qui-quadrado, maior e mais significativa a diferença entre o modelo restrito e o modelo completo.

O teste que compara o modelo proposto a um modelo perfeitamente ajustado é dado por:

$$n \ln \left\| \Sigma(\theta) \right\| + \text{tr} S \Sigma^{-1}(\theta) - n \ln \left\| \Sigma(\hat{\theta}) \right\| - p. \quad (17)$$

Caso o teste não seja significativo, não podemos rejeitar a hipótese de o modelo proposto não ser diferente do modelo perfeitamente ajustado.

Uma vantagem dos testes de RV é que são úteis em situações onde existem complicações paramétricas, isto é, parâmetros que estão presentes no modelo, mas não são de interesse direto em termos inferenciais, por outro lado, uma desvantagem do teste de RV é que requer estimadores de MV dos parâmetros sob o modelo geral e sob a hipótese nula, para o qual pode necessitar de um processo iterativo, uma vez que, em geral, não é possível obter expressões explícitas para estas estimativas sob a hipótese nula (Oliveira, 2001).

Para que melhor possa avaliar a qualidade do teste de ajuste, definimos as hipóteses nula e alternativa respectivamente como  $H_0: \Sigma = \Sigma(\theta)$  versus  $H_1: \Sigma$  é uma matriz de covariância sem restrição. Seja a matriz de covariância residual populacional definida por  $E_p$  de tal forma que  $E_p = \Sigma - \Sigma(\theta)$ . Quando  $H_0$  é verdadeiro,  $E_p$  é uma matriz de zeros. A matriz de covariância residual amostral  $\hat{E}$ , definida como  $\hat{E} = [s - \Sigma(\theta)]$ , que poderão nos auxiliar na identificação da matriz de componentes de variância-covariância que não estão bem ajustados.

Neste projeto pretendemos avaliar as diferentes deficiências contidas no bloco deficiência em relação às variáveis dos outros blocos de variáveis que diz respeito a identificação, família, trabalho e instrução mencionados no tópico descrição das variáveis.

### 12.3 - Estimação e ajuste

A avaliação do ajuste do modelo pode ser feita em duas partes: o exame do ajuste local e do ajuste global. A avaliação do ajuste local é feita através da estatística do teste "t". Deste modo pode ser verificado se os parâmetros são significativamente diferentes de zero. Informação sobre o ajuste local também pode ser obtida, analisando o coeficiente de determinação  $R^2$  e os resíduos. A avaliação do ajuste global do modelo pode ser feita pelo teste de ajuste exato, usando o teste qui-quadrado, e pelos testes de ajuste aproximado usando RMSEA e RMR. Outros critérios de ajustes aproximados estão disponíveis como o GFI, CFI e NFI (Gutierrez, 2006).

A estimação dos parâmetros se baseia na seleção dos valores dos parâmetros estruturais que reproduzem a matriz de covariância.

Os parâmetros do modelo fatorial confirmatório podem ser estimados usando-se a máxima verossimilhança com a seguinte função objetiva:

$$\ln L = \frac{n}{2} \left( \ln \|\Sigma(\theta)\| + tr(S\Sigma^{-1}(\theta)) \right) \quad (18)$$



onde  $\ln(L)$  é o log de verossimilhança,  $S$  é a matriz de covariância,  $\Sigma$  uma função de parâmetros do modelo e  $n$  é o número de observações.

No MEE, podemos estimar os valores do parâmetro do modelo para maximizar a verossimilhança da matriz de covariância observada, no qual o modelo da verdadeira matriz de covariância Sigma conforme expressão (6).

O procedimento de estimativa por MV fornece estimativas de erro padrão assintótico de cada parâmetro do modelo, que podem ser usados para realizar teste  $t$  de significância.

Os testes qui-quadrados de ajuste de modelo e os índices de qualidade de ajuste utilizados em AFC, podem também, ser utilizados para avaliar MEE com variáveis latentes.

É importante verificar se o método de estimativa convergiu de fato para uma solução viável. Certifique-se de que não haja variância de erros negativas e que todas as estimativas de parâmetros de outro modelo façam sentido.

Entre os possíveis critérios destacam-se inicialmente Função de discrepância é uma função matemática que descreve o quanto um modelo estrutural está fechado com relação aos dados observados. Grandes valores para esta função indicam um pior ajuste dos dados. Esta função foi desenvolvida por Jöreskog (1967) a partir da MV e é dada por

$$F_{MV} = \ln|\Sigma(\theta)| + \text{tr}(S\Sigma^{-1}(\theta)) - \ln|S| - p \quad (19)$$

onde  $p$  é o número de variáveis que se manifesta no modelo e  $\Sigma(\theta)$  é a matriz de covariância populacional do modelo proposto. Utilizando a suposição de normalidade multivariada para os dados observados produz assintoticamente os erros padrões corretos para as estimativas dos parâmetros e um completo ajuste estatístico que segue, assintoticamente uma distribuição qui-quadrado quando o modelo está correto na população.

Temos que os estimadores de MV possui propriedades importantes como: embora possam ser viciados, eles são assintoticamente não viciados; são consistentes; são assintoticamente eficientes; e por fim, tem distribuição assintoticamente normal.

Alternativamente, a função de discrepância por mínimos quadrados generalizados(GLS) é dada por:

$$F_{GLS} = \text{tr}(I - S\Sigma^{-1}(\theta))^2 \quad (20)$$

Esta função produz resultados que são assintoticamente equivalente aos resultados da mesma função utilizando estimativas por MV. Conforme Walt David tem destacado, uma

diferença significativa entre as duas funções de discrepância que está no escore que ela associa ao modelo "nulo", que é, modelos que especificam todas as variáveis que se manifesta que são não correlacionadas. Os valores das funções de discrepância por estimativas de MV e por estimativas de mínimos quadrados generalizados podem variar muito. Os resultados que são índices de ajustes comparativos baseados no modelo nulo não comportarão de forma consistente através destes dois métodos de estimação.

As funções de estimação de MV ou GLS podem ser utilizadas para o cálculo de estimativas dos parâmetros do modelo quando a distribuição das variáveis observadas desvia-se da normal, porém, os erros padrões e as medidas de qualidade de ajuste qui-quadrado devem ser interpretados com cautela (Melhado, 2004).

A função de ajuste para o método de mínimos quadrados não ponderados (ULS) é dada por:

$$F_{ULS} = \frac{1}{2} \text{tr} \left[ \left( S - \Sigma^{-1}(\theta) \right)^2 \right] \quad (21)$$

Após o ajuste do modelo é necessário analisar rigorosamente os coeficientes estimados e a formação de associações. Ou seja, devemos verificar se os sinais dos coeficientes estão de acordo com a realidade que se deseja traduzir pelo modelo, isto é, se a correlação entre as variáveis é realmente positiva, negativa ou nula; e, os parâmetros estimados são estatisticamente significantes.

Para realizar a validação do modelo, existem algumas medidas de ajustes que são apresentadas a seguir.

**Critério ICSF** mede quanto o modelo estrutural é invariante sob um fator de constante escalar, se o ajuste do modelo não for alterado quando todas as variáveis são multiplicadas pela mesma constante, e por fim;

**Critério ICS** representa o quanto o modelo estrutural é invariante sob mudança de escala se o ajuste do modelo não for mudado pelo recálculo das variáveis, isto é, pela multiplicação delas pelos fatores de escala.

O ajuste do modelo pode ser avaliado em duas partes: o exame do ajuste local e o exame do ajuste global. A avaliação do ajuste local é feita através da estatística de teste *t*. Deste modo pode ser verificado se os parâmetros são significativamente diferente de zero. A avaliação do ajuste global do modelo, pode ser feita pelo teste de ajuste exato, usando a

estatística de teste qui-quadrado, e pelo teste de ajuste aproximado usando a estatística RMSEA, a raiz do erro quadrático médio (Gutierrez, 2006).

A estatística qui-quadrado é a RV que se utiliza para avaliar a significância estatística de que todos os elementos da matriz de resíduos sejam nulos.

A **Raiz do Erro Quadrático Médio (RMSEA)**, proposto por Staiger e Lind (1989), é a raiz quadrada da média dos resíduos quadrados ajustados a qual só pode ser interpretada em relação aos tamanhos das variâncias e covariâncias observadas em S. Esta medida é mais útil quando todas as variáveis observadas estão padronizadas. Não existe ponto de corte a partir do qual se possa concluir que o ajuste seja aceitável. Parâmetro de não centralidade populacional é estimado diretamente.

A raiz do erro quadrático médio leva em consideração o erro de aproximação na população, isto é, verifica quanto o modelo se ajusta à matriz de covariâncias populacional caso esta estivesse disponível (com parâmetros desconhecidos, mas com a escolha ótima dos parâmetros). O RMSEA mede esta discrepância ou erro. Seu cálculo é realizado da seguinte maneira:

$$RMSEA = \sqrt{\frac{\hat{F}_0}{gl}}, \quad (22)$$

em que  $\hat{F}_0 = \max\left(\frac{\hat{C} - gl}{N}; 0\right)$ . Esta medida leva em consideração o erro de aproximação na população. É possível, também, utilizar intervalos de confiança para avaliar à precisão da estimativa de RMSEA.

Os limites inferior e superior do intervalo de confiança de 90% ( $L_{0.90}$  e  $H_{1.90}$ , respectivamente).

Os limites são dados por:  $L_{0.90} = \sqrt{\frac{\delta_L/N}{gl}}$  e  $H_{1.90} = \sqrt{\frac{\delta_V/N}{gl}}$  em que  $\delta_L$  e  $\delta_V$  são obtidos, respectivamente, pela solução das equações:  $\chi^2_{nc}(C/\delta, gl) = 0,95$  e  $\chi^2_{nc}(C/\delta, gl) = 0,05$ . Onde  $\chi^2_{nc}(C/\delta, gl)$  é a função de distribuição qui-quadrado não central com parâmetro de não centralidade  $\delta$  e  $gl$  graus de liberdade. Browne and Cudeck (1993) indicam que valores de RMSEA menores que 0,08 é um bom ajuste para o modelo.

O RMSEA tem como vantagens representar melhor o quão bem um modelo se ajusta a uma população e não apenas a uma amostra usada para estimação e a sua principal vantagem

é que um intervalo de confiança pode ser construído, fornecendo o domínio RMSEA para um dado nível de confiança.

**Índice de não centralidade de McDonald**, proposto por McDonald (1989), representa uma aproximação para transformação do índice de não centralidade populacional  $F^*$  no intervalo entre 0 e 1. Este índice é calculado por  $\exp\left(-\frac{F^*}{2}\right)$ .

Uma vantagem é que este índice representa uma transformação para o índice  $F^*$  de não centralidade populacional para um intervalo entre 0 e 1.

**Índice de população Gamma** é um índice que tenta corrigir o tamanho amostral e a complexidade do modelo pela inclusão de cada uma dessas informações em seus cálculos. Sua principal vantagem é que ela tem uma distribuição conhecida, e pode ser calculado da seguinte maneira:

$$\Gamma_1 = \frac{p}{2F^* + p} \quad (23)$$

em que  $p$  é o número de variáveis que se manifesta e  $F^*$  é o índice de não centralidade da população e uma desvantagem é que, desta maneira,  $\Gamma_1$ , proporciona a adição de um maior número de parâmetros.

**Índice de população Gamma ajustado**, neste caso, penaliza os modelos mais complexos subtraindo  $p^* = \frac{p(p+1)}{2}$ , em que  $p$  é o número de parâmetros do modelo e o índice é dado por:

$$\Gamma_2 = 1 - (p^*/\nu)(1 - \Gamma_1) \quad (24)$$

Como o teste qui-quadrado é altamente influenciado pelo tamanho da amostra (isto é, quanto maior  $n$ , maior será a tendência de julgar todos modelos como inadequados), os pesquisadores desenvolveram outros índices como GFI e AGFI descritos a seguir:

**Índice de Qualidade de Ajuste (GFI)**, proposto por Jöreskog and Sörbom (1986), representa um grau de ajuste global (resíduos quadrados dos dados estimados em relação aos dados observados) não envolvendo os graus de liberdade. O intervalo possível dos valores GFI é entre 0 e 1, e, quanto maior este índice, melhor deverá ser o ajuste, que pelo método de MV pode ser calculado por:

$$GFI_{MV} = \frac{1 - \text{tr}\left[\left(\Sigma^{-1}(\theta)S - I\right)^2\right]}{\text{tr}\left[\left(\Sigma^{-1}(\theta)S\right)^2\right]} \quad (25)$$

Aplicando o método de mínimos quadrados não ponderados (ULS), o GFI pode ser calculado como:

$$GFI_{ULS} = 1 - \frac{P(P+1)}{2gl} \frac{tr\left[\left(\Sigma^{-1}(\theta)S - I\right)^2\right]}{tr\left[\left(\Sigma^{-1}(\theta)S\right)^2\right]}. \quad (26)$$

**Índice de Qualidade de Ajuste corrigido (AGFI)** tenta levar em conta diferentes graus de complexidades do modelo. Ele faz isso ajustando GFI por uma proporção entre os graus de liberdade usados em um modelo e o número total de graus de liberdades disponíveis. O AGFI penaliza modelos mais complexos e favorece aqueles com um número mínimo de caminhos livres. Valores AGFI são tipicamente menores que valores GFI em proporção a complexidade do modelo, que pode ser calculado por

$$AGFI_{MV} = 1 - \frac{P(P+1)}{2gl} \frac{tr\left[\left(\Sigma^{-1}(\theta)S - I\right)^2\right]}{tr\left[\left(\Sigma^{-1}(\theta)S\right)^2\right]}. \quad (27)$$

Aplicando método ULS o AGFI pode ser calculado por:

$$AGFI_{ULS} = 1 - \frac{P(P+1)}{2gl} \frac{tr\left[\left(\Sigma^{-1}(\theta)S - I\right)^2\right]}{tr\left[\left(\Sigma^{-1}(\theta)S\right)^2\right]}. \quad (28)$$

Aplicando o método de mínimos quadrados generalizados (GLS), proposto por Tanaka e Huba(1985), o GFI pode ser calculado como:

$$GFI_{GLS} = 1 - \frac{tr\left[\left(I - \Sigma^{-1}(\theta)S^{-1}\right)^2\right]}{p}. \quad (29)$$

Já, para o AGFI, pode ser calculado por:

$$AGFI_{GLS} = 1 - \frac{p(p+1)}{2gl} \frac{tr\left[\left(I - \Sigma^{-1}(\theta)S^{-1}\right)^2\right]}{p}. \quad (30)$$

Para os casos de  $GFI_{MV}$  e  $AGFI_{MV}$  em estudo de simulações feito por Anderson e Gerbing (1984), eles obtiveram que: as médias de distribuições amostrais de  $GFI_{MV}$  e  $AGFI_{MV}$  aumentaram com o aumento do tamanho da amostra; e,  $GFI_{MV}$  e  $AGFI_{MV}$  decresceram de acordo com o aumento do número de indicadores por fator ou número de fatores.

Segundo Lattin et al., (2011) é considerado ajuste forte para um determinado modelo quando  $GFI > 0,95$  e  $AGFI > 0,90$ .

O AGFI procura levar em conta diferentes graus de complexidade do modelo, ajustando o GFI por uma proporção entre os graus de liberdade usados em um modelo e o número total de graus de liberdade disponíveis, com isso, o AGFI penaliza modelos mais complexos e favorece aqueles com um número mínimo de caminhos livres.

O GFI mede a quantidade relativa das variâncias e covariâncias em S que é predita por  $\hat{\Sigma}$  e, AGFI corrige GFI de um modelo pelos graus de liberdade relativos ao número de variáveis.

A vantagem do GFI e AGFI é que são dimensionados entre zero e 1, o que facilita a comparação entre modelos alternativos, e não é uma função do tamanho da amostra, porém, estes critérios possuem como desvantagens que não há um padrão objetivo (tal como um teste estatístico) que estipule o que representa um nível aceitável de GFI ou AGFI, porém, sabe-se que quanto mais próximo de 1, maior será a probabilidade do ajuste do modelo ser considerado aceitável, e, permanece sensível de forma indireta ao tamanho da amostra  $n$  sobre distribuições amostrais (Maiti and Mukherjee, 1991).

**Índice Comparativo Ajustado** é dado por

$$CFI = 1 - \frac{\max(\hat{C}_m - gl_m, 0)}{\max(\hat{C}_b - gl_b, 0)} \quad (31)$$

em que  $\hat{C}_m$  e  $\hat{C}_b$  são as discrepâncias mínimas da amostra do modelo corrente e do modelo "baseline", respectivamente e,  $gl_m$  e  $gl_b$  são seus respectivos graus de liberdade. CFI varia entre 0 e 1 e, quanto mais próximo de 1, mais aceitável é o modelo;

**Índice normalizado ajustado**, proposto por Bentler e Bonett (1980), é dado por

$$\Delta_1 = \frac{F_b - F_m}{F_b} = \frac{\chi_b^2 - \chi_m^2}{\chi_b^2} \quad (32)$$

em que  $F_b$  é o valor da função de ajuste do modelo "baseline" e  $F_m$  é o valor da função de ajuste do modelo "corrente". Ele mede a redução na função de ajuste quando passa do modelo "baseline" para o modelo "corrente".  $\Delta_1$  assume valores entre 0 e 1, quanto mais próximo de 1, melhor será o ajuste.

Este índice possui algumas limitações como:

O valor de  $F_m$  pode ser reduzido pela adição de parâmetros, aumentando o valor de  $\Delta_1$  (análogo ao aumento de  $R^2$  na equação de regressão quando adiciona uma variável

explicativa); o tamanho da amostra pode influenciar a média da distribuição amostral de  $\Delta_1$ , pois:

$$(n-1)F_m \propto \chi_{g_m}^2 \Rightarrow F_m \propto \frac{\chi_{g_m}^2}{n-1} \Rightarrow E(F_m) \approx \frac{g_m}{n-1}. \quad (33)$$

Note que quanto maior  $n$ , menor  $E(F_m)$  e, o mesmo acontece com  $F_b$ , porém, mais lentamente que  $E(F_m)$ . Assim, comparando duas amostras de tamanhos diferentes podemos concluir que a amostra de tamanho maior tem um melhor ajuste mesmo se os modelos idênticos forem ajustados para ambas as amostras.

**Índice de Ajuste não normalizado ajustado** é calculado por

$$\Delta_2 = \frac{F_s - F_m}{F_s - \left[ \frac{g_m}{(N-1)} \right]} = \frac{\chi_s^2 - \chi_m^2}{\chi_s^2 - g_m}. \quad (34)$$

O índice de ajuste não normalizado corrigido tem as seguintes propriedades para um modelo corretamente ajustado: não varia necessariamente entre 0 e 1; valores muito acima de 1 podem ser devidos ao superajuste (pode ocorrer se pelo menos uma das condições for verdadeira: i) o número de variáveis é maior que o número de observações; e, as variáveis são altamente correlacionadas ou multicolinear, (Searle, 2004)); por fim, quando o numerador e o denominador de  $\Delta_2$  são positivos,  $\Delta_1 < \Delta_2$ . Porém, observe-se que a média da distribuição amostral de  $\Delta_2$  deve estar menos relacionada ao tamanho da amostra do que a média da distribuição amostral de  $\Delta_1$ ; o cálculo de  $\Delta_2$  é influenciado pelo tamanho da amostra, pois, quanto menor o tamanho da amostra, maior será o valor de  $\Delta_2$ ; e por fim, quanto maior o tamanho da amostra, os valores de  $\Delta_2$  e  $\Delta_1$  ficam próximos.

**Índice de Ajuste parcimonioso**, proposto por Mulaik et al., (1989) é o índice corrigido pela razão entre os graus de liberdade do modelo "corrente" e o número de parâmetros do "modelo saturado", isto é,  $PGFI = GFI \frac{2gl}{k(k-1)}$ . Este índice varia entre 0 e 1. Um modelo é considerado parcimonioso quando possui boa capacidade explicativa e não contém parâmetros desnecessários. Trata-se de um índice utilizado para comparar 2 modelos;

**Bollen's Rho**, proposto por Bollen (1989), é um índice de ajuste comparativo que calcula a redução relativa da função de discrepância pelo grau de liberdade quando varia do modelo nulo para o modelo  $k$ , e é calculado por:

$$\rho_k = \frac{\frac{F_0 - F_k}{V_0 - V_k}}{\frac{F_0}{V_0}}. \quad (35)$$

onde  $\nu_0$  é o número de graus de liberdade para o modelo nulo;  $\nu_k$  é o número de graus de liberdade para o modelo  $k$ ;  $F_0$  é a função de discrepância para o modelo nulo; e por fim,  $F_k$  é a função de discrepância para o modelo  $k$ , observe que a adição de parâmetros diminui os graus de liberdade, logo, é possível ver que  $\rho_k$  se mantenha constante ou diminua; modelos ajustados que relativamente tem poucos parâmetros e um valor baixo da função de ajuste possuem valores menores para  $\rho_k$  de que modelos com o mesmo valor da função de ajuste e um modelo mais complexo; se  $F_m \rightarrow 0 \Leftrightarrow \rho_k \rightarrow 1$ ; é possível que  $\frac{\chi_m^2}{gl_m}$  seja maior que  $\frac{\chi_k^2}{gl_k}$  levando a  $\rho_k < 0$ .

Uma desvantagem na utilização desta medida é que  $\rho_k$  aumenta proporcionalmente ao tamanho da amostra  $n$ , pois  $(n-1)F_m \alpha \chi_{gl_m}^2 \Rightarrow \frac{F_m}{gl_m} \alpha \chi_1^2 \Rightarrow \frac{F_m}{gl_m} \alpha \chi_{\frac{1}{n-1}}^2 \Rightarrow E\left(\frac{F_m}{gl_m}\right) \alpha \frac{1}{n-1}$

**Bollen's Delta** apenas faz a correção do modelo de Bollen's Rho para simplificar nos casos em que ocorre alto grau de liberdade e é dado por:

$$\Delta_k = \frac{F_0 - F_k}{F_0 - \left[\frac{\nu_k}{n}\right]}; \quad (36)$$

onde  $F_0$ ,  $F_k$  e  $\nu_k$  são os mesmos definidos para Bollen's Rho e  $n$  é o tamanho da amostra.

Qui-quadrado é uma medida que é calculada por:

$$\chi_p^2 = \frac{\chi^2 - gl}{\sqrt{2 \cdot gl_m}}; \quad (37)$$

O  $\chi_p^2$  mede o desvio entre a estatística qui-quadrado e seu valor esperado em termos do desvio padrão da estatística  $\chi^2$ .

**Qui-quadrado relativo** é um índice obtido dividindo o valor da estatística qui-quadrado pelos seus graus de liberdade. Isto é,  $\chi^2/gl$ . O qui-quadrado relativo mede quantas vezes a estatística qui-quadrado ( $\chi^2$ ) é maior que seu valor esperado ( $gl$ ).

Uma vantagem ao utilizar o teste qui-quadrado é que o tamanho amostral não afeta os graus de liberdade, mas por outro lado, o teste qui-quadrado apresenta duas propriedades matemáticas que são problemáticas em seu emprego que são a estatística qui-quadrado que depende do tamanho da amostra  $n$  e da diferença entre as matrizes de covariância observada e estimada; e também, a estatística qui-quadrado pode aumentar quando o número de variáveis aumenta.



**Critério de Informação de Akaike (AIC)** é um índice alternativo para comparar modelos baseados em Akaike (1974) e é dado por:

$$AIC = \hat{C} + 2t, \quad (38)$$

em que  $t$  é o número de parâmetros livres e  $\hat{C}$  é a discrepância mínima da amostra para o modelo proposto. Recomenda-se selecionar o modelo com menor valor de AIC (Bollen, 1989), sendo significativa as diferenças de AIC maiores que 2.

Segundo Sakamoto et al. (1986), AIC é um critério geral corrigido pelo número de parâmetros a ser incluído no modelo.

Para Nakamichi et al. (2001), o critério AIC é escolhido como critério de ajuste por apresentar alto poder preditivo sobre a distribuição dos dados do modelo. A maior vantagem de utilizar o AIC neste caso é excluir variáveis cujos efeitos não seriam omitidos por outras estratégias de análise. Pela maximização do poder de predição, o AIC cria um balanceamento entre os erros tipo I e II, e, neste sentido é desejável aplicá-lo na predição de variáveis do modelo.

**Critério de Informação Bayesiano (BIC):** é um índice dado por:

$$BIC = \hat{C} + t \times \ln\left(\frac{N}{k}\right), \quad (39)$$

em que  $\hat{C}$  e  $t$  são os mesmos descritos anteriormente,  $N$  é o número de casos e,  $k$  é o número de variáveis observadas.

Este índice oferece uma penalidade maior para a complexidade do modelo e possui uma maior tendência em escolher modelos mais parcimoniosos (Arbuckle, 1997). Deve-se escolher o modelo com menor valor de BIC.

A principal diferença entre AIC e o BIC é que o BIC admite que o modelo que gerou os dados, ou seja, o modelo "verdadeiro", se encontra entre o conjunto de modelos candidatos e procura encontrá-lo, enquanto o AIC busca encontrar, dentre os modelos disponíveis, o modelo mais adequado (Oliveira, 2008).

Com base no exposto, na comparação entre modelos com o mesmo número de parâmetros, o critério SSE, além de ser de fácil implementação, assegura um bom desempenho do procedimento de seleção de modelos, pois consegue discriminar bem a qualidade dos ajustes entre dois ou mais modelos competitivos (Carlborg et al., 2000; Wang, 2000). Este critério é considerado neste projeto por assegurar uma melhor adequabilidade e discriminação

dos diferentes modelos sob estudo.

Por outro lado, como alternativa ao teste da RV para a comparação simultânea de diferentes modelos existem diversas medidas baseadas nos critérios de informação. Como visto, de maneira geral, estas são calculadas com um termo de penalização aplicado à função de verossimilhança, as mais comuns sendo os critérios de Akaike (AIC) e o critério de Schwartz (BIC). Em adição ao SSE, os critérios AIC e BIC são também utilizados neste projeto para seleção de modelos.

A penalização considerada para o critério AIC foi  $2n_{par}$  e para o critério BIC a penalização foi  $-n_{par}\ln(n)$ , com a justificativa de facilitarem o cálculo de máximo por simulações (Paulino et al., 2003) e de serem consideradas boas opções de penalização em cada caso (Wang et al., 2007).

**Fator de Bayes:** Em um contexto Bayesiano, dois modelos competitivos podem ser comparados usando o Fator de Bayes  $FB$ , obtido pela razão das evidências a posteriori e a priori de dois modelos. Considere a comparação entre os modelos  $Mod_1$  e  $Mod_2$ . O  $FB$  é dado por: (Paulino et al., 2003)

$$FB_{Mod_1,Mod_2} = \frac{p(dados|Mod_1)}{p(dados|Mod_2)}; \quad (40)$$

onde o numerador é o *odds* a posteriori,  $p(Mod_1|dados)p(Mod_2|dados)$ , e o denominador é o *odds* a priori,  $p(dados|Mod_1)p(dados|Mod_2)$ .

O Fator de Bayes avalia diretamente a superioridade de um modelo sobre um outro (Kass and Raftery, 1995). Quando as probabilidades a priori dos dois modelos são as mesmas, isto é, não existe por parte do pesquisador, um favoritismo a priori para um ou outro modelo antes da realização do experimento, o  $FB$  é igual ao *odds* a posteriori (Sillampää and Corander, 2000).

Jeffreys (1935) foi o primeiro a utilizar o Fator de Bayes na comparação de diferentes teorias científicas, criando inclusive uma regra para tomada de decisões em função de seu valor. A Tabela 1 apresenta uma calibração para o Fator de Bayes baseado no valor de  $FB$ .

Tabela 1: Decisões sobre evidência de $Mod_1$ em relação a $Mod_2$	
Valores de $FB$	Conclusão
$1 \leq FB \leq 3$	Evidência a favor de $Mod_1$
$3 < FB \leq 10$	Evidência positiva a favor de $Mod_1$
$10 < FB \leq 100$	Forte evidência a favor de $Mod_1$
$FB > 100$	Evidência decisiva a favor de $Mod_1$

**Fonte:** Jeffreys (1935) e Raftery (1995)

É importante observar que em situações que envolvem modelos com muitos parâmetros, o cálculo algébrico do Fator de Bayes torna-se inviável (Toledo, 2006), envolvendo a avaliação de integrais de alta dimensão. Para estas situações muitas aproximações têm sido propostas, entre as quais o BIC (do inglês, *Bayesian Information Criterion*) é uma alternativa mais simples e precisa na classe de distribuições regulares (Schwartz, 1978; Raftery, 1995).

Critério de Browne-Cudeck (BCC): é um critério dado por

$$BCC = \hat{C} + 2t \frac{b \frac{k(k+3)}{N-k-2}}{k(k+3)}; \quad (41)$$

em que  $\hat{C}$ ,  $t$ ,  $N$  e  $k$  são os mesmos descritos anteriormente e  $b = N-1$ . Deve-se escolher o modelo com menor valor de BCC. Em geral, melhores modelos terão menores índice para BCC.

**Soma de Quadrados dos Resíduos (SSE):** A soma de quadrados de resíduos (SSE) é um critério de seleção de modelos que se refere diretamente à adequabilidade do modelo. Pode ser calculada para modelos que incluem muitas variáveis independentes. Este valor é dado por:

$$SSE = Y' [I - X(X'X)^{-1} X'] Y; \quad (42)$$

onde  $Y$  e  $X$  estão definidos no modelo proposto. Para seleção de modelos é considerado "melhor" o modelo que tiver menor SSE (Wang, 2000; Carlborg, 2000; Kutner et al., 2005).

**Quadrado Médio do Resíduo (MSE):** O método do quadrado médio residual (MSE), ou erro quadrático médio (Kutner et al., 2005) é um critério de seleção de modelos que, tal como o SSE, se refere à adequabilidade do modelo. Inicialmente é calculado o MSE, como dado na expressão

$$MSE = \frac{SSE}{n - n_{par}} = \frac{(X - Y\hat{\beta})'(X - Y\hat{\beta})}{n - n_{par}} = \frac{Y'[I - X(X'X)^{-1} X'] Y}{n - n_{par}}; \quad (43)$$

para os diferentes modelos competitivos. Note que esta medida padroniza o SSE pelo seu grau de liberdade. Quando modelos com diferentes valores de  $n_{par}$  estão sendo comparados, a medida de ajuste MSE é mais indicada comparada ao SSE. Calcula-se então  $\bar{s}^2$ , que é a média dos valores MSE obtidos para todos os modelos com o mesmo número de parâmetros  $n_{par}$ . A seguir é construída a curva com os valores de  $\bar{s}^2$  em função de  $n_{par}$ . À medida que variáveis predictoras desnecessárias são acrescentadas,  $\bar{s}^2$  tende a se estabilizar (Zeng, 2001; Kutner et

al., 2005). O modelo selecionado é aquele para o qual MSE é próximo do valor de  $\bar{s}^2$  que estabilizou, contendo número mínimo de parâmetros (isto é, com  $n_{par}$  mínimo).

**Critério  $C_p$ :** O critério  $C_p$  assume a seguinte fórmula (Mallows, 1973; Miller, 1990):

$$MSE = \frac{SSE}{\hat{\sigma}^2} - (n - 2n_{par}); \quad (44)$$

onde  $\hat{\sigma}^2$  é um estimador de  $\sigma^2$ , por exemplo, como dado na expressão (34). O modelo selecionado é aquele que minimiza este critério. Note que, para que a equação (35), com  $n_{par}$  parâmetros, seja adequada como medida de ajuste,  $\hat{\sigma}^2$  deve ser um estimador não viciado de  $\sigma^2$  e, assim, o valor médio de  $C_p$  deve ser próximo de  $n_{par}$ . Os casos em que  $C_p > n_{par}$  correspondem a  $E(SSE) > (n - n_{par})\hat{\sigma}^2$ . Assim, segundo esse critério, será escolhido o modelo com valor de  $C_p$  baixo, próximo de  $n_{par}$  e com uma quantidade mínima de variáveis preditoras. Tendo em vista estas considerações, o critério  $C_p$  leva à seleção de um modelo com valor de  $C_p$  o mais próximo possível de  $n_{par}$ . Este critério recebe críticas (Miller, 1990), pois, em geral, o modelo selecionado envolve um número muito pequeno de parâmetros, e modelos de maior ordem (com mais variáveis preditoras) podem estar sendo descartados mesmo tendo ajustes melhores.

**Critério  $R^2$ :** O critério de maior  $R^2$  também pode possibilitar a escolha do "melhor" modelo (Schuster and Cruz, 2004; Euclides Filho, 2000). A medida  $R^2$ , também conhecida como coeficiente de determinação, é definida como:

$$R^2 = \frac{SSR}{SST}; \quad (45)$$

onde SSR e SST são as somas de quadrados do modelo de regressão e total, respectivamente. O problema deste critério é que o modelo com maior  $R^2$  é sempre o que contém um maior número de variáveis regressoras quando comparado com um modelo reduzido. Porém, muitas vezes, o aumento em  $R^2$  com a inclusão de mais variáveis preditoras no modelo é tão pequeno, que não justifica o modelo mais complexo. Desta forma, uma possível análise sistemática é dividir as equações de regressão em  $k$  grupos de acordo com o número de variáveis preditoras e selecionar, dentro de cada grupo, os modelos com maior  $R^2$ . Este procedimento permite estudar o comportamento de  $R^2$  com o acréscimo de variáveis preditoras no modelo e verificar se um possível aumento de  $R^2$  é suficiente de modo a justificar a adoção de um modelo com maior quantidade de variáveis. Além da complexidade do modelo, outra desvantagem de utilizar muitas variáveis preditoras é o possível aparecimento de multicolinearidades. Existem

disponíveis na literatura muitas alternativas de se trabalhar com este efeito (Kutner et al., 2005; Draper and Smith, 1998).

**Critério  $R^2$  ajustado:** Como uma forma de estabelecer uma correção ao critério  $R^2$ , Miller (1990) propôs a seguinte alternativa:

$$R^2 \text{ ajustado} = 1 - (1 - R^2) \frac{n}{n - n_{par}}; \quad (46)$$

onde  $R^2$  ajustado é o coeficiente de determinação corrigido pelo tamanho amostral  $n$  e os graus de liberdade do resíduo ( $n - n_{par}$ ). Observe que o  $R^2$  ajustado utiliza o número de parâmetros como uma forma de penalização. O modelo selecionado neste caso será aquele que tiver o maior valor de  $R^2$  ajustado entre os modelos candidatos. Um problema apontado na literatura é que os critérios SSE e  $R^2$  ajustado tendem a incluírem uma grande quantidade de variáveis desnecessárias, o que torna estes métodos mais liberais ao serem comparados com o método  $C_p$ .

**Estatística da Razão de Verossimilhanças:** Considerando a comparação entre modelos encaixados, isto é, que se diferenciam apenas pela inclusão ou não de um conjunto de parâmetros (em geral, associados a variáveis preditoras), uma alternativa é adotar a estatística RV ( $-2\ln(LR)$ ), que requer a avaliação do ajuste do modelo sob duas hipóteses, nula e alternativa. Uma formulação da estatística da RV é apresentada em

$$A = -2\ln(LR) = -2\ln\left(\frac{L_0}{L_1}\right). \quad (47)$$

O modelo selecionado será aquele com o qual os dados amostrais são mais verossímeis, o que é avaliado pelo valor da estatística razão de verossimilhanças comparado com um correspondente valor crítico. Limitações de uso deste critério se refere à comparação entre modelos não encaixados, envolvendo diferentes variáveis preditoras, por exemplo.

Uma outra possível medida de ajuste para quantificar a possível diferença entre as matrizes de covariância  $S$  e  $\Sigma(\theta)$  é a matriz de resíduos ( $Y_{(k \times k)}$ ). Os elementos dessa matriz são  $v_{ij} = s_{ij} - \hat{\sigma}_{ij}$ , onde  $s_{ij}$  é o  $ij$ -ésimo elemento de  $S$  e  $\hat{\sigma}_{ij}$  é o elemento correspondente em  $\Sigma(\theta)$ .

É possível afirmar que um modelo é "bom" se todos os resíduos estiverem próximos de zero (Melhado, 2004).

Uma medida alternativa, baseada nos resíduos amostrais  $v_{ij}$  é a raiz do quadrado médio residual (RMR), que é dada por:

$$RMR = \left[ 2 \sum_{i=1}^k \sum_{j=1}^i \frac{v_{ij}^2}{k(k+1)} \right]^{\frac{1}{2}} = \left[ 2 \sum_{i=1}^k \sum_{j=1}^i \frac{(s_{ij} - \hat{\sigma}_{ij})^2}{k(k+1)} \right]^{\frac{1}{2}} \quad (48)$$

É considerado um bom ajuste quando RMR é próximo de zero. Mas, os resíduos amostrais podem ser afetados por fatores como: distância entre  $\Sigma$  e  $\Sigma(\theta)$  e, por escalas diferentes entre as variáveis observáveis. Esta medida pode ser adotada para comparar o ajuste entre dois diferentes modelos para o mesmo conjunto de dados. Além disso, de acordo com (Jöreskog and Sörbom, 1989; Jöreskog and Sörbom, 1997), esta medida geralmente conduz a conclusões muito similares quando comparadas ao valor da média absoluta dos resíduos não padronizados.

Esta medida é útil para comparar ajustes ao longo de modelos e permite ao pesquisador localizar problemas potenciais com um modelo de mensuração e uma desvantagem é que nenhum valor estatístico pode ser estabelecido.

Resíduos amostrais não são apenas afetados pela diferença entre  $\Sigma$  e  $\Sigma(\theta)$ , mas também, pelas escalas de  $Y_i$  e pelos erros. Assim, uma medida alternativa para solucionar este problema é utilizar o resíduo da correlação ( $l_{ij}$ ).

$$l_{ij} = r_{ij} - \hat{r}_{ij}; \quad (49)$$

em que:  $r_{ij}$  é a correlação amostral entre a  $i$ -ésima e  $j$ -ésima variáveis;  $\hat{r}_{ij}$  é a correlação entre a  $i$ -ésima e a  $j$ -ésima variáveis no modelo predito, e que é dado por.

$$\hat{r}_{ij} = \frac{\sum(\hat{\theta})_{ij}}{\sqrt{\sum(\hat{\theta})_{ii} \sum(\hat{\theta})_{jj}}}; \quad (50)$$

Como  $-1 \leq r_{ij}, \hat{r}_{ij} \leq 1$ , temos que  $-2 \leq l_{ij} \leq 2$ . Valores de  $l_{ij}$  próximos de zero indicam um bom ajuste e próximo de -2 indicam problemas no ajuste.

Através dos resíduos amostrais ( $l_{ij}$ ), é possível dizer que: se  $l_{ij} > 0$ , o modelo subestima a covariância entre as duas variáveis, caso contrário, o modelo superestima a covariância.

O processo de estimação do modelo consiste das seguintes etapas: especificação; estimação; avaliação e, às vezes, mais uma; modificação do modelo. Na Figura 5 está apresentado as etapas deste procedimento (Gutierrez, 2006).

Nota-se na Figura 5 um resumo do procedimento de cada uma das etapas da estimação de um Modelo de Equações Estruturais.

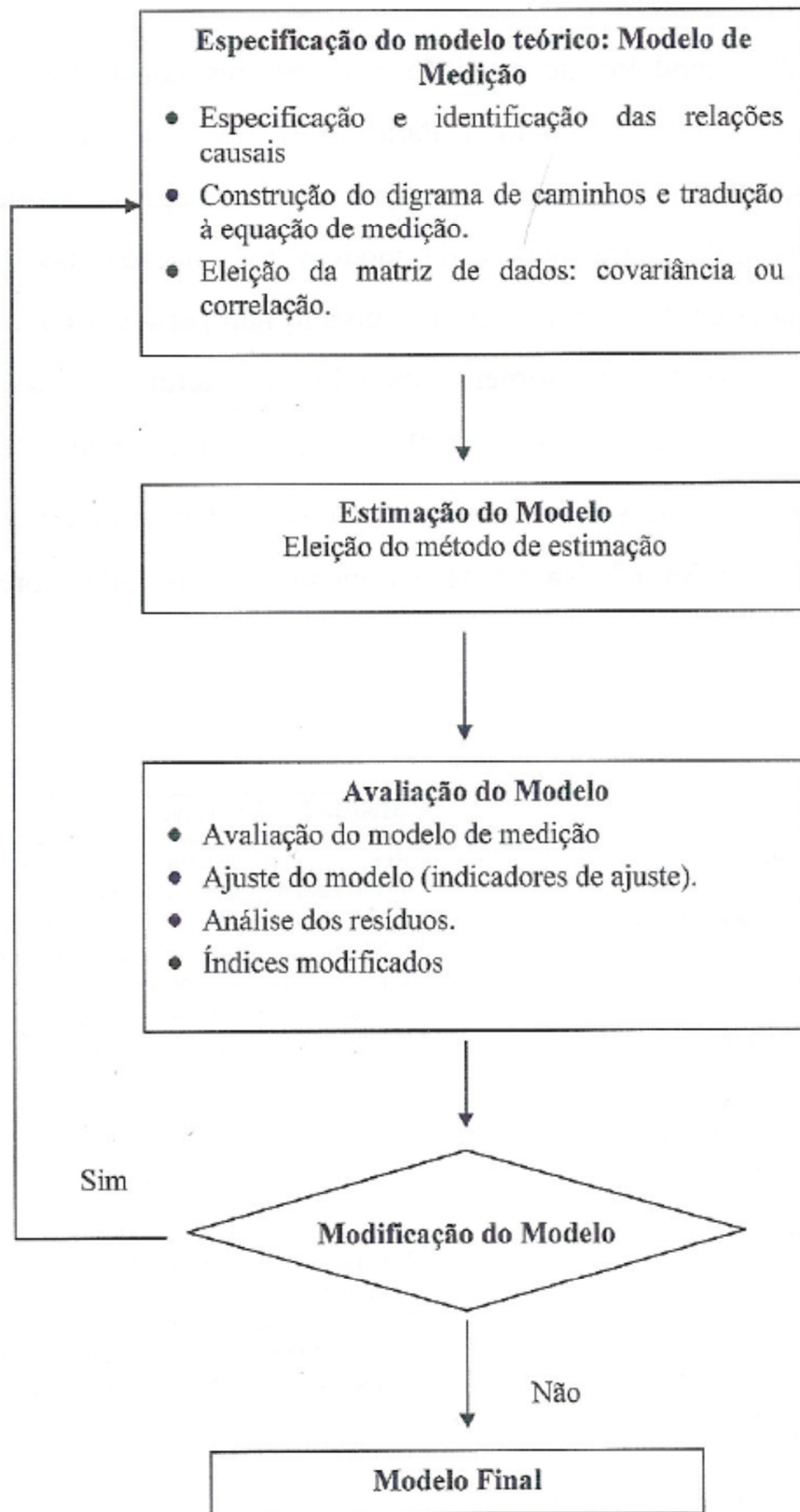


Figura 5. Algoritmo de procedimento de estimação para Modelo de Equações Estruturais.

#### **12.4 – Suposições violadas e suas consequências**

Conforme já mencionada, uma das suposições do modelo de equações estruturais é a de que as variáveis observadas são contínuas. Entretanto, podem existir situações em que faz necessária a incorporação de variáveis categóricas ordinais e nominais que sejam relevantes para se estimar os construtos do modelo.

Ao adotar este tipo de procedimento, o pesquisador pode estar tendo um problema de perda de confiabilidade nos resultados invalidando qualquer interpretação obtida.

Em vista disso, nota-se a necessidade de examinar quais suposições são violadas com incorporação de indicadores categóricos, avaliando suas consequências, e, apresentar alguns dos procedimentos corretivos (Bistaffa, 2010).

No trabalho de Bistaffa (2010) a discussão se restringiu à discussão de variáveis latentes contínuas com indicadores dicotômicos ou ordinais e sua extensão a modelos de equações estruturais incluindo propostas de procedimentos de correção para estes casos (Lazarsfeld and Henry, 1968; Goodman, 1972; Bishop et al., 1975; Ameniya, 1981; Maddala, 1983).

Outra alternativa a ser considerada como possibilidade de correção das violações do modelo que pode ser causada devido a incorporação de variáveis categóricas é a utilização de outros métodos de estimação ou até mesmo de outras distribuições com caudas mais leves ou mais pesadas do que a normal como as distribuições elípticas (Melhado, 2009).

No caso deste projeto, pretendemos estender essa discussão considerando relações mais complexas como também a incorporação de variáveis categóricas nominais como sexo, raça e estado civil.

#### **12.5 – Modelos elípticos**

A suposição usual de MEE é a normalidade multivariada, porém, quando esta suposição é violada, os testes de hipóteses correspondentes podem oferecer resultados enganosos. Neste caso, uma alternativa para corrigir esse tipo de problema é a utilização de distribuições elípticas que possui uma vasta gama de distribuições com coeficiente de assimetria 0, muito embora possa ter curtose diferente da curtose da distribuição normal multivariadas.

A fórmula geral de uma densidade elíptica é dada por:



$$c|V|^{-1/2}h\{(z-\mu)'V^{-1}(z-\mu)\} \quad (51)$$

sendo que  $c$  é uma constante;  $h$  é uma função não-negativa, e, por fim;  $\mathbf{z}$  é um vetor de média  $\boldsymbol{\mu}$  e matriz positiva-definida  $\mathbf{V}$ .

Neste caso a função a ser ajustada se resume a:

$$F_E = \frac{1}{2}(c+1)^{-1}tr\left[(S-\Sigma(\Theta))V^{-1}\right]^2 - c_1tr\left[(S-\Sigma(\Theta))V^{-1}\right]^2, \quad (52)$$

sendo que  $c$  é um parâmetro de curtose comum;  $c_1 = \left[\frac{c}{4(c+1)^2 + 2(p+q)k(k+1)}\right]$ , e, por fim;  $\mathbf{V}$  é uma matriz de pesos.

O parâmetro de curtose comum pode ser estimado por:

$$\hat{c} = \frac{b_{2(p+q)} - (p+q)(p+q+2)}{(p+q)(p+q+2)} \dots \quad (53)$$

sendo que  $b_{2(p+q)} = \left(\frac{1}{n}\right)\left\{(Z_i - \bar{Z})'S^{-1}(Z_i - \bar{Z})\right\}^2$ .

O parâmetro de curtose comum também pode ser estimado por:

$$\hat{c} = \frac{1}{3(p+q)} \sum_{i=1}^{p+q} b_{2i}, \quad (54)$$

sendo que  $b_2$  são as estimativas de curtose univariadas.

Escolhendo  $\mathbf{V}$  como um estimador consistente de  $\Sigma$  obtemos um estimador assintoticamente eficiente de  $\Theta$  e  $(n-1)F_E$  possui distribuição assintoticamente  $\chi^2$ . A escolha mais comum de  $\mathbf{V}$  é  $\mathbf{S}$  (Bollen, 1989). Neste projeto, pretendemos testar a validade do ajuste por diferentes modelos elípticos como alternativa para correção da violação de suposições devido a utilização de dados categóricos no modelo de MEE.

## 12.6 - Seleção de variáveis

Perante um modelo de regressão com um elevado número de preditores, o objetivo será reter apenas algumas dessas variáveis e eliminar as restantes do modelo. Existindo um grande número de preditores pode-se ter problemas entre estas variáveis, tal como a multicolinearidade. Também é desejável selecionar um conjunto menor de variáveis que ajuste tão bem quanto o modelo com todas as variáveis e que contenha as variáveis predictoras mais importantes (Kyung et al, 2010).

Entre as diferentes estratégias que podem ser utilizadas para selecionar variáveis destacamos forward stepwise e a backward stepwise. A forward stepwise inicia com a constante  $\beta_0$ , e sequencialmente adiciona ao modelo o preditor  $X_i$  mais correlacionado com  $Y$  de modo que melhore o ajuste de acordo com a avaliação da estatística  $F$ , e a introdução de variáveis para quando não consegue produzir uma estatística  $F$  maior de que o percentil 90 ou 95 da distribuição  $F_{1, N-k-2}$ , em que  $N$  é o tamanho amostral e  $k$  é o número de variáveis.

Já, a estratégia de seleção backward stepwise inicia com o modelo com todas as variáveis independentes, e sequencialmente, vai excluindo variáveis usando a estatística  $F$  para escolher os preditores a eliminar. O preditor que possui a menor estatística  $F$  é eliminado e o processo para quando cada preditor eliminado do modelo tem um valor  $F$  maior que o percentil de 90 ou 95 da distribuição  $F_{1, N-k-2}$ ,

Após os parâmetros terem sido estimados, o próximo passo é avaliar se as covariáveis utilizadas e disponíveis para a modelagem são estatisticamente significantes com o evento modelado, como por exemplo, a condição de um indivíduo ser pessoa com deficiência.

Uma forma de testar a significância do coeficiente de uma determinada covariável é comparar os valores observados da variável resposta com os valores preditos obtidos pelos modelos com e sem a variável de interesse (Oliveira, 2008).

A comparação entre os valores observados e preditos é feito a partir do teste de razão de verossimilhanças que é largamente aplicável com a estimação de máxima verossimilhança.

Para testar  $H_0 : \theta \in \Theta_0$  versus  $H_a : \theta \in \Theta_0^c$  calculamos a estatística (Casella and Berger, 1990):

$$\lambda(x) = \frac{\sup_{\theta_0} L(\theta/x)}{\sup_{\theta} L(\theta/x)}. \text{ Para } n \rightarrow \infty, -2 \ln \lambda(x) \rightarrow \chi^2_{\nu}.$$

onde,  $\nu$  é obtido através da diferença entre a quantidade de parâmetros existente no modelo testado e a quantidade de parâmetros existente no modelo saturado (Oliveira, 2001).

A qualidade do modelo ajustado pode ser verificado através da comparação entre os valores observados e os valores preditos para a variável resposta (neste caso pode ser uma das diferentes deficiência já mencionadas).

Na escolha de um particular modelo, se por um lado, devemos tentar incluir tantas variáveis independentes quanto possíveis para melhorar a previsão, por outro, desejamos incluir um número mínimo de variáveis por problemas de custo e simplicidade (Oliveira, 2008).

Segundo Draper e Smith (1998) define seleção do melhor modelo como o compromisso de conciliar estes dois objetivos. Esta seleção envolve uma dose de subjetividade o resultado pode ser diferente se alterar o procedimento utilizado para seleção.

Neste projeto, pretendemos implementar, para cada um dos diferentes modelos métodos de seleção de variáveis considerando os diferentes critérios de ajuste.

### **12.7 - Seleção de modelos**

Selecionar um modelo significa, após a formulação e ajuste de diferentes modelos plausíveis, selecionar o modelo que "melhor" se ajusta aos dados de um certo experimento de acordo com um determinado critério estabelecido (Camarinha Filho, 2002). Fisher (1922) discutiu três aspectos do problema geral de inferência: especificação, estimação dos parâmetros e precisão da estimativa. Na estimação ou ajuste do modelo, bem como na precisão de estimativas. Para o MEE, a especificação do modelo leva em conta o número de parâmetros e quais seus efeitos. Na estimação ou ajuste do modelo e também, pode-se adotar metodologias paramétricas ou não paramétricas, sendo as primeiras mais comuns (Oliveira, 2008).

O problema de identificação de modelos é que deve-se formar um critério para comparação entre os diferentes modelos. Em geral, a seleção de modelos é justificada para explicar, prever ou simplificar. Em explicação deseja modelos que tenham boas aproximação ao "modelo verdadeiro". Em predição deseja modelos com boa qualidade de prever e em simplificação deseja a facilidade de interpretação. Este problema significa que deve-se formar um critério para comparação entre os diferentes modelos. Modelos com o mesmo número de parâmetros são geralmente comparados pelo critério da soma de quadrados dos resíduos (SSE) obtidos após a estimação dos coeficientes de regressão por mínimos quadrados, por exemplo. Sob este critério, o modelo com a menor valor da soma de quadrados dos resíduos é selecionado. Quando se compara modelos com diferentes números de parâmetros, não se pode simplesmente selecionar o modelo com o menor valor para a soma de quadrados dos resíduos, pois desde que um novo parâmetro (variável preditora) seja adicionado ao modelo, a soma de quadrados dos resíduos tende a diminuir. Nestes casos, penalizações devem ser incluídas nos critérios.

A necessidade de selecionar modelos surge quando o pesquisador necessita decidir qual dos modelos utilizar quando tem em mãos um certo número de modelos equivalentes levando-se em conta apenas o conhecimento subjetivo a priori. Um princípio bastante utilizado é o princípio da parcimônia, isto é, o menor número possível de parâmetros para uma

adequada representação dos dados (Burnham and Anderson, 1998). Em geral, conforme aumenta a dimensão  $k$  do modelo o viés diminui e a variância aumenta. Dessa forma, um modelo é parcimonioso se for tão simples quanto possível com respeito as variáveis incluídas, estrutura do modelo e de parâmetros.

Existem vários procedimentos para a seleção de modelos, embora, nenhum deles seja consistente, ou seja, mesmo para amostras grandes selecione com probabilidade um as variáveis explicativas com coeficiente não nulo (Paula, 2013).

A seleção de modelos, segundo Broman (2001) exige quatro passos distintos: i) Selecionar uma classe de modelos: refere-se à especificação do modelo; ii) Pesquisar diretamente o espaço dos possíveis modelos e dar ênfase, quando de interesse, a modelos particulares; iii) Adotar um critério de seleção e comparar os diferentes modelos relativamente a este critério; e, por fim, iv) Assegurar o desempenho do procedimento, isto é, assegurar que o valor do critério atribuído a cada modelo mensure seu grau de aptidão ou adequabilidade e que o conjunto destes valores gere uma escala que diferencie entre os modelos.

Por outro lado, uma medida de sucesso de seleção de modelos depende da frequência com que o verdadeiro modelo é selecionado.

Em Estatística há uma vasta literatura pertinente à seleção de modelos (Broman, 1997; Burnham and Anderson, 1998, 2002) e a utilização de modelos baseados na função de verossimilhança e os critérios de Akaike (Paulino et al., 2003; Sakamoto et al., 1986). Outras propostas estão na classe de metodologia bayesiana, como o critério de Schwartz ou fator de Bayes e suas modificações (Broman, 1997; Wang, 2000).

Na estrutura Bayesiana, duas hipóteses concorrentes podem ser comparadas usando o FB, que proporciona uma avaliação direta de superioridade de um modelo sobre o outro (Paulino et al., 2003). Quando as probabilidades à priori são iguais, o FB é igual à chance das posteriores. A avaliação do FB pode envolver a avaliação de integrais de alta dimensão que é muitas vezes intratável. Diversas alternativas ao fator de Bayes têm sido propostas, entre as quais, destaca-se o BIC, que é igualmente simples e preciso quando usado em modelos regulares, portanto, a aproximação pelo BIC é geralmente apropriada e inferências Bayesianas podem ser feitas com um mínimo adicional em termos computacionais (Oliveira and Pavan, 2005).

A dificuldade nestes casos é a comparação de modelos que não necessariamente são encaixados (sequência de modelos, cujos modelos reduzidos com menor quantidade de

parâmetros, são subconjuntos dos modelos saturados que apresentam maior quantidade de parâmetros).

A necessidade de selecionar modelos surge quando um pesquisador deseja, entre diferentes modelos, que o modelo escolhido tenha uma boa aproximação do "modelo verdadeiro" (explicação); boa qualidade de previsão e, por fim, facilidade de interpretação (explicação).

Vários procedimentos de seleção de modelos e critérios estatísticos podem ser considerados para auxiliar na elaboração de decisões sobre os termos que devem ser incluídos ou excluídos do modelo. Em geral, a seleção de modelos está ligada à estimação dos parâmetros associados ao modelo estatístico adotado ou à predição do comportamento de variáveis aleatórias envolvidas na definição do modelo (Burnham and Anderson, 1998, 2002). Metodologias, como SSE, AIC e BIC, são bastante flexíveis (Blangero et al., 2001) e podem ser utilizadas com vantagem se acopladas a algoritmos computacionais como modelos lineares (Oliveira, 2008).

### **12.8 – Simulações**

Pretendemos realizar simulações para que possam testar a robustez avaliando assimetria e curtose, avaliar o comportamento e comparar as diferenças entre as medidas de qualidade de ajuste e também nos valores dos parâmetros estimados em cada caso para diferentes tamanhos amostrais utilizando matriz de correlação de Pearson e a matriz híbrida de correlações especiais quando o modelo é estimado pelas diferentes funções de discrepância considerando variáveis categóricas ordinais, nominais e contínuas (Melhado, 2009; Bistaffa, 2010; Brunelli, 2012).

## **13 - Geoestatística**

Geoestatística é um ramo da Geografia matemática e da estatística que une o conceito de variáveis aleatórias com o conceito de variáveis regionalizadas, gerando um novo conceito de funções aleatórias, que são posteriormente processadas por aplicativos computacionais. Através destas técnicas, dentre as quais se destacam a krigagem e a simulação estocástica, é possível calcular um valor de uma dada propriedade (permeabilidade, porosidade, etc.) para cada centro da célula de uma malha tridimensional, valor este condicionado aos dados existentes (dados de poços, sísmica, residência de uma pessoa com deficiência, etc.) e a uma função de correlação espacial entre estes dados.

Em várias áreas das Ciências da Terra, as variáveis não apresentam um padrão de distribuição requerido pela estatística clássica como normalidade e independência dos dados. Os modelos da estatística clássica estão geralmente voltados para a verificação da distribuição de frequência dos dados, enquanto a **geoestatística** incorpora a interpretação da distribuição estatística, assim como a correlação espacial das amostras. Este aspecto da geoestatística está intimamente associado com a distribuição estatística dos dados no espaço (Diggle and Ribeiro Jr., 2007).

Assim, os métodos geoestatísticos fornecem um conjunto de ferramentas para entender a uma aparente aleatoriedade dos dados, mas com possível estruturação espacial, estabelecendo, desse modo, uma função de correlação espacial. Esta função representa a base da estimativa da variabilidade espacial em geoestatística.

Um exemplo de como um método adequado de representar dados, aqui eminentemente espacial, pode indicar a solução de um problema complexo, mesmo sem conhecer de forma precisa as causas do fenômeno em análise encontra-se bem evidente no trabalho de John Snow (1855) sobre a epidemia da cólera, que alastrou em Londres no século XIX.

A 31 de agosto de 1854 surgiu em Londres uma epidemia de cólera que matou mais de 500 pessoas em cerca de 10 dias. O alarme foi geral e os londrinos abandonaram a cidade em massa, o que provocou uma forte interrupção da vida econômica e social da cidade.

O gráfico da Figura 6 representa o número de mortes por cólera registradas diariamente durante a epidemia, mostra bem o rápido desenvolvimento da doença.

No entanto, J. Snow estava convencido que a doença era transmitida através da água contaminada e, para tentar provar essa teoria, registrou no mapa da parte de Londres onde a epidemia alastrava, a localização precisa da moradia de cada um dos 83 mortos por cólera de uma lista que obteve do General Register Office.

O mapa que desenhou, do qual apresentamos na Figura 7 uma cópia, mostra claramente que a maioria dos mortos, assinaladas com pequenos traços negros nas respectivas residências, se verificaram nas imediações do poço situado na confluência da Broad Street com a Cambridge Street marcado com dois círculos no mapa.

Repare-se na Figura 7 que o desenho de um único círculo significa poço não afetado, dois círculos é o poço infectado da Broad Street e o retângulo preta representa a quantidade de mortes por cólera no local, proporcional ao comprimento do mesmo.

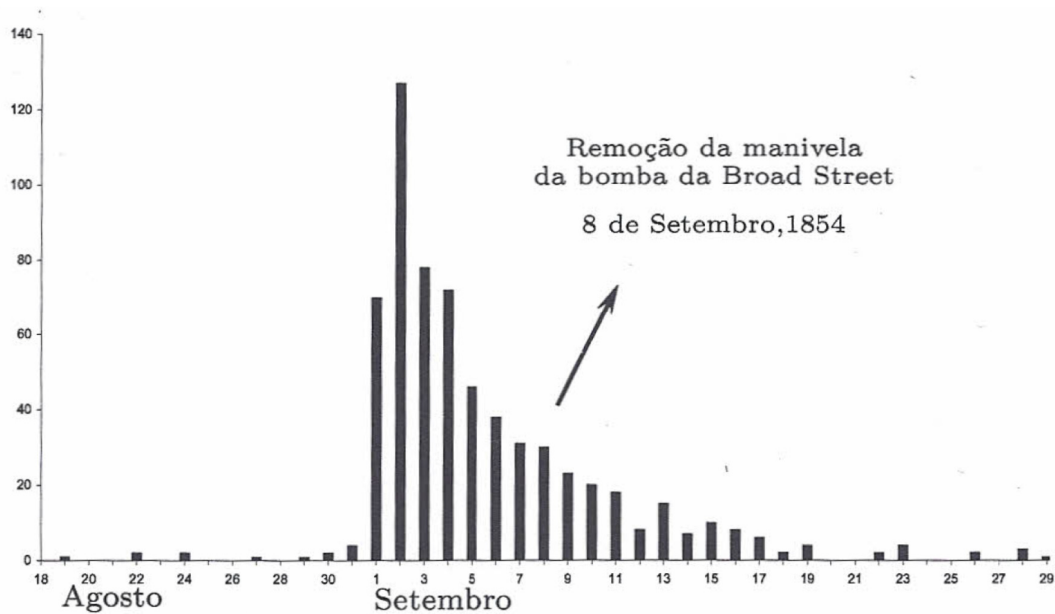


Figura 6. Óbitos por cólera em cada dia da epidemia.

Apesar do mapa ser bastante informativo, Jonh Snow fez uma investigação porta a porta sobre os hábitos de abastecimento de água dos 83 dos casos, onde pode confirmar a correlação existente entre as mortes e a utilização de água do poço da Broad Street.

Embora tivesse observado que a água do poço não apresentava impurezas suspeitas. Snow descreveu as suas descobertas às autoridades responsáveis pelo abastecimento de água a 7 de setembro e, no dia seguinte, a manivela da bomba do referido poço foi removida.

É erradamente, atribuída à remoção do manípulo o fim deste surto de cólera. De fato, se verificamos na Figura 6, a epidemia nessa data, já se encontrava na sua fase descendente, muito provavelmente devido à fuga do local de grande parte da população. No entanto, a não utilização do poço contaminado foi certamente responsável pela prevenção de novos surtos quando os ocupantes regressaram às suas residências.

Em 1886 a descoberta do vibrião do cólera veio confirmar a hipótese tão brilhantemente avançada por Snow, cujo o trabalho é com frequência apontado como o primeiro exemplo de aplicação da representação espacial de dados na descoberta de solução para um problema de saúde pública.

No caso aqui, este método pode ser aplicado para determinar a distribuição e predição ao longo do país das pessoas com as deficiências em função de outras variáveis como nível de instrução, renda, sexo, raça, condições de moradia e outras.



Figura 7. Mapa de Londres onde deflagrou a epidemia de cólera.

Para este projeto, estamos considerando que as residências das pessoas com deficiência foram divulgadas até a localização de subdistrito, transformando este estudo numa análise de dados areolares. De fato, os dados referentes a residência das pessoas com deficiência envolve confidencialidade, e, como tal, só podem ser divulgados de forma agrupada sem risco de perda de anonimato.

Os mapas como os das figuras 1 e 2 são conhecidos como mapas de manchas, pelo fato de utilizarem manchas de cor para classificar as variáveis em estudo para classificar os valores das variáveis em estudo, por exemplo, existência de pelo menos uma deficiência na Figura 1, eu um pequeno número de classes. Estes mapas possuem como vantagem de fornecer uma ideia sumária, facilmente interpretável da distribuição regional da característica de interesse. A desvantagem desta simplicidade é a falta de robustez da mensagem que transmitem, que se apoia essencialmente na sua natureza visual, facilmente manipulável por alteração dos limites das classes ou do gradiente de cores utilizado.



No presente projeto estamos interessados em saber se existirá alguma tendência para os casos de pessoas com deficiência se aglomerarem em certas zonas do país, após controlado os efeitos que a concentração populacional nas zonas litorâneas e a acentuada e acentuada diferença entre a distribuição das pessoas com deficiência dessas zonas e do interior ao produzirem o mapa apresentado.

A adoção de modelos que permitem construir estimativas alisadas do número de pessoas com deficiência em cada região possibilita a obtenção de imagens mais realistas da incidência de deficiência.

Para este trabalho, adotaremos técnicas de análise espacial de dados agregados por áreas passando pelas seguintes etapas:

- a) Considerar como modelo inferencial a ser utilizado, a hipótese de que as áreas são diferenciadas, e que cada uma delas possui “identidade” própria, isto é, sob o ponto de vista estatístico, implica que em cada área apresenta uma distribuição de probabilidade distinta das demais, ou seja, modelo espacial discreto.
- b) Propomos como modelo de distribuição, o modelo de variação espacial discreta. Neste caso, consideraremos a existência de um processo estocástico  $z_i$ ,  $i = 1, \dots, n$ , onde  $z_i$  é a realização do processo espacial na área  $i$  e  $n$  é o total de áreas  $A_i$ . O objetivo principal da análise é construir uma aproximação para a distribuição conjunta de variáveis aleatórias  $Z = \{z_1, \dots, z_n\}$  estimando sua distribuição, neste caso,  $z_i$  representa a quantidade de pessoas com deficiência em uma área de tamanho  $i$ .

Neste caso,  $z_i$  é a variável aleatória que descreve a contagem, indicador ou taxa associada à área  $A_i$ . Dispomos de um valor observado  $z_i$  correspondente à contagem na  $i$ -ésima área que descreve o número de ocorrências em cada área associada a uma distribuição de Poisson.

- c) Análise exploratória

Para que possa ocorrer um melhor desenvolvimento das etapas de modelagem estatística espacial, utilizaremos técnicas de análise exploratórias sensíveis ao tipo de distribuição; à presença de valores extremos utilizando ferramentas gráficas como histograma e box plots para que possa investigar outliers no conjunto de dados e nos vizinhos; e, por fim, à ausência de estacionariedade estudando variação na média (primeira ordem), na variância e na correlação espacial.

Neste projeto, consideraremos como análise exploratória a utilização de técnicas para visualização de dados, entre as quais destacando-se:

- Visualização dos dados que possibilita a visualização de valores extremos no mapa, considerar diferentes métodos de corte de variáveis como intervalos iguais, percentis e desvios padrões;
- Gráficos de médias e medianas segundo linhas e colunas, para que possa explorar simultaneamente a presença de tendência (não estacionariedade de primeira ordem) e não estacionariedade de segunda ordem, onde a variância e covariância entre vizinhos não se mantem constantes. Para construir estes gráficos, utilizaremos as coordenadas dos centroides de áreas e calcularemos as médias e as medianas do indicador (eixo Leste-Oeste) e coluna (eixo Norte-Sul). Esta técnica permite identificar a ação das medidas ao longo das duas flutuações, sugerindo a presença de valores discrepantes quando a diferença entre estas é grande e a tendência ao longo de uma direção quando os valores variam suavemente.

Na Figura 8, apresenta-se o resultado desta técnica aplicadas a dois indicadores socioeconômicos do censo de 1991, renda média do chefe de família e proporção de chefes de famílias com escolaridade igual ou superior ao segundo grau, para setores censitários da Ilha de Governador, no Rio de Janeiro. Esta é composta por 225 setores censitários, cujos centroides estão assinalados no primeiro quadro da figura: observe que nas extremidades do “mapa” a quantidade de pontos é muito pequena, e, conseqüentemente, qualquer medida nesta área será pouco robusta.

Verifica-se na Figura 8 que no eixo Norte-Sul (colunas) pode-se observar que a renda média do chefe da família apresenta tendência variável, bem menor no centro da região. A mesma coisa acontece para escolaridade, em bora com maior flutuação. No eixo Leste-Oeste (linhas), também parece haver algum deslocamento para valores mais altos no sentido leste, mas o deslocamento de médias e medianas sugere a presença de valores extremos dos indicadores. A variação na média dos indicadores na região está aparentemente, dividida entre as duas direções analisadas, e pode-se explorar melhor a tendência através da rotação dos eixos de referência.

- Análise de auto correlação espacial que melhor descreva os dados. A ideia básica é estimar a magnitude da auto correlação espacial entre as áreas. Neste caso, consideraremos o índice global de Moran, o índice de Geary e o variograma;

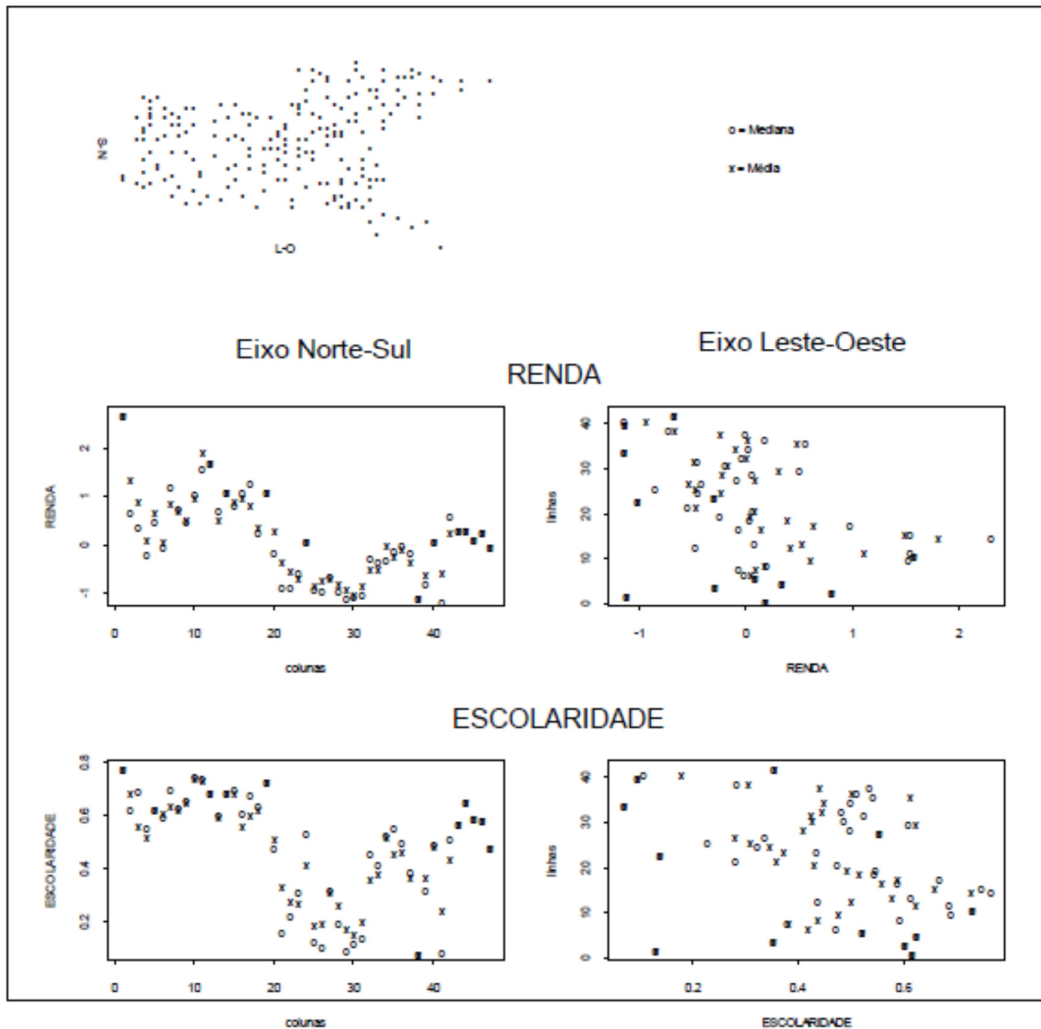


Figura 8. Médias e medianas para escolaridade e renda na Ilha de Governador.

- Matrizes de proximidade espacial ou matriz de vizinhança, com a finalidade de estimar a variabilidade espacial de dados de área. Dada um conjunto de  $n$  áreas  $(A_1, \dots, A_n)$ , construímos a matriz  $W^{(1)}$  ( $n \times n$ ), onde cada um dos elementos  $w_{ij}$  representa uma medida de proximidade entre  $A_i$  e  $A_j$ . Esta medida de proximidade pode ser calculada a partir de um dos seguintes critérios:
  - $w_{ij} = 1$ , se o centroide de  $A_i$  está a uma determinada distância de  $A_j$ , caso contrário,  $w_{ij} = 0$ ;
  - $w_{ij} = 1$ , se  $A_i$  compartilha um lado comum com  $A_j$ , caso contrário  $w_{ij} = 0$ ;
  - $w_{ij} = l_{ij}/l_i$ , onde  $l_{ij}$  é o comprimento da fronteira entre  $A_i$  e  $A_j$  e  $l_i$  é o perímetro de  $A_i$ .

A Figura 9 ilustra um exemplo de matriz de proximidade espacial, em que os valores dos elementos da matriz refletem o critério de adjacência

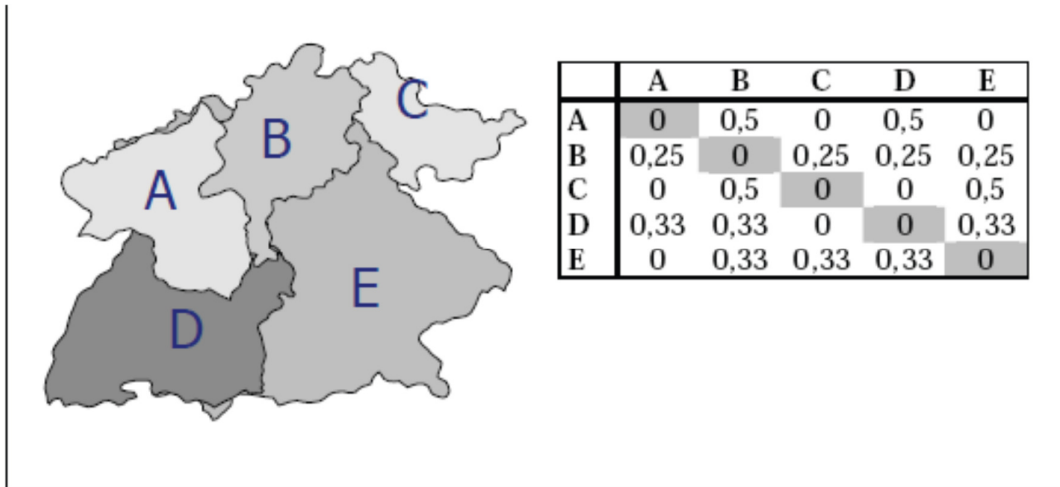


Figura 9. Matriz de proximidade espacial de primeira ordem, normalizada pelas linhas.

- Média móvel espacial ou média dos valores dos vizinhos, para que possa explorar a variação da tendência espacial dos dados. Isto reduz a variabilidade espacial, pois a operação tende a produzir uma superfície com menor flutuação que os dados originais. A média móvel  $\hat{\mu}_i$  associada ao atributo  $z_i$ , relativo a  $i$ -ésima área, pode ser calculada a partir dos elementos  $w_{ij}$  da matriz normalizada de proximidade espacial  $W^{(0)}$ , tornando-se simplesmente a média dos vizinhos:

$$\hat{\mu}_i = \sum_{j=1}^n w_{ij} z_j \quad (55)$$

A Figura 10 ilustra o uso do estimador de média móvel para o percentual de idosos (acima de 70 anos) para os 96 distritos da cidade de São Paulo. Estes dados são indicadores de grande disparidade social da cidade com uma grande variação entre o centro (proporção de idosos chegando a 8%) com a periferia (com essa proporção chegando a menos de 1%).

Pode-se notar ao comparar os dois mapas da Figura 8 que a média móvel local fornece uma visão das grandes tendências do percentual idoso, mostrando um forte gradiente entre centro periferia.

- Indicadores globais de auto correlação espacial  
Entre os diversos índices de correlação espacial existentes destacamos os índices de Moran e de Geary.

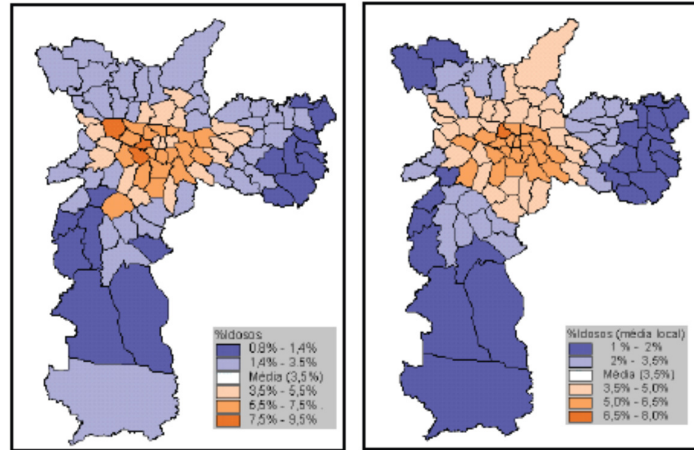


Figura 10. Distribuição dos idosos na cidade de São Paulo (censo de 1991) a) apresentação dos valores por distribuição estatística. b) média móvel local.

Para que possa caracterizar a dependência espacial, mostrando como os valores estão correlacionados no espaço. Neste contexto, utilizaremos as funções para estimar o quanto o valor observado de um atributo numa região é dependente dos valores desta mesma região é dependente dos valores desta mesma variável nas localizações vizinhas são a auto correlação espacial e o variograma.

Trata-se o índice global de Moran a expressão de auto correlação considerando apenas o primeiro vizinho

$$I = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (z_i - \bar{z})(z_j - \bar{z})}{\sum_{j=1}^n (z_i - \bar{z})^2} \quad (56)$$

Na equação (56)  $I$  é o índice global de Moran;  $n$  é o número de áreas;  $z_i$  é o valor do atributo na área  $i$ ;  $\bar{z}$  é o valor médio do atributo na região de estudo e  $w_{ij}$  são os elementos da matriz normalizada de proximidade espacial. O cálculo feito para matrizes de proximidade espacial. O cálculo feito para matrizes de maior ordem permite estimar a função de auto correlação para cada ordem de vizinhança (ou "lag").

$$I^{(k)} = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij}^{(k)} (z_i - \bar{z})(z_j - \bar{z})}{\sum_{j=1}^n (z_i - \bar{z})^2} \quad (57)$$

De forma geral, o índice de Moran presta-se a um teste cuja hipótese nula é de independência espacial, e, a seguir, estabelece a significância

estatística verificando-se estes valores de auto correlação está associada a uma determinada distribuição.

A hipótese implícita no cálculo do índice de Moran é a estacionariedade de primeira e segunda ordem, e este índice perde sua validade ao ser calculado para dados não estacionários. Quando existir não estacionariedade de primeira ordem (tendência), os vizinhos tenderão a ter valores mais parecidos que de áreas distantes. Da mesma forma, se a variância não for constante, nos locais de maior variância o índice será mais baixo e vice-versa. Algumas variações deste modelo são o  $C$  de Geary e o teste  $I_{pop}$ . O primeiro ( $C$  de Geary) difere do teste  $I$  de Moran por utilizar a diferença entre os pares, enquanto que Moran utiliza a diferença entre cada ponto e a média global. Assim, o indicador  $C$  de Geary assemelha-se ao variograma e o  $I$  de Moran ao correlograma

$$C = \frac{(n-1) \sum_{i=1}^n \sum_{j=1}^n w_{ij} (z_i - z_j)^2}{\sum_{j=1}^n w_{ij} (z_i - \bar{z})^2} \quad (58)$$

O teste  $I_{pop}$ , trata-se de outro teste utilizado para detectar desvios de uma distribuição aleatória, porém, incorpora a variação da população nas áreas. Assim, torna-se sensível à ocorrência de aglomerados intra-área e entre-áreas. O índice  $I_{pop}$  pode ser decomposto em uma componente intra-área e outra entre áreas, que podem ser apresentada sob forma percentual nos resultados. A hipótese nula assume que a variação geográfica do tamanho da população, sendo útil quando a população das áreas não é homogênea.

$$I_{pop} = \frac{N^2 \sum_{i=1}^m \sum_{j=1}^m w_{ij} (e_i - d_i)(e_j - d_j) - N(1 - 2\bar{b}) \sum_{j=1}^m w_{ij} e_i - N\bar{b} \sum_{j=1}^m w_{ij} d_i}{X^2 \left( \sum_{i=1}^m \sum_{j=1}^m d_i d_j w_{ij} - X \sum_{j=1}^m d_i w_{ij} \right) \bar{b} (1 - \bar{b})} \quad (59)$$

- Variograma: trata-se de outra técnica que pode ser utilizada como indicador da dependência espacial. Para tanto, associamos o valor único do atributo de cada área ao ponto representado pelo centro geométrico ou populacional do polígono. É importante notar também que quando o dado não é estacionário, também o variograma não se estabiliza. Como exemplo de uso do variograma

para dados de área, a Figura 9 ilustra o Índice de desenvolvimento humano (IDH) para o Estado de São Paulo, com base no censo de 1991. A Figura 11 representa o variograma do IDH calculado a partir do centroide de cada município.

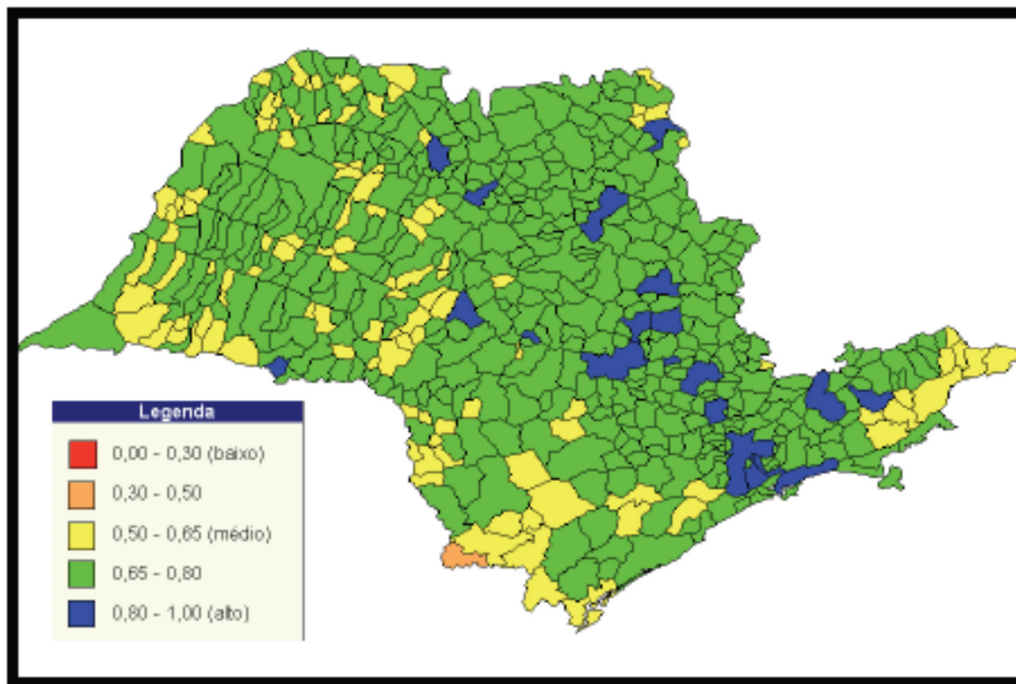


Figura 11. IDH para São Paulo (censo de 1991).

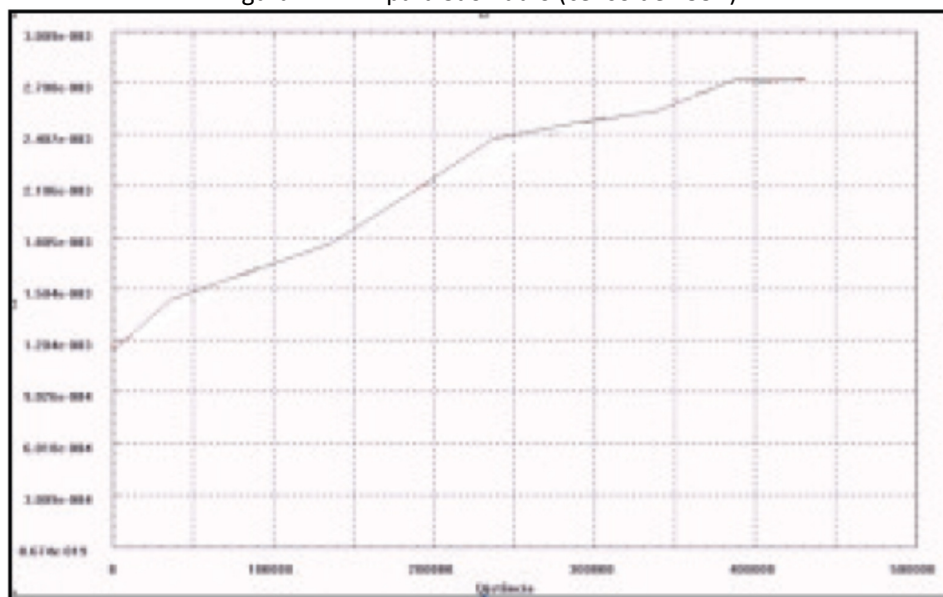


Figura 12. Variograma experimental do IDH para São Paulo (censo de 1991). Passo de amostragem 40km (tolerância: 20km).

- Diagrama de espalhamento de Moran: Trata-se de uma maneira adicional de visualizar a dependência espacial. Construído com base nos valores

normalizados, permite analisar o comportamento da variabilidade espacial. A ideia é comparar os valores normalizados do atributo numa área com a média de seus vizinhos, construindo um gráfico bidimensional de valores normalizados pela média dos vizinhos, conforme mostrado na Figura 13 para o índice de inclusão/exclusão de São Paulo, censo de 1991. Os quadrantes podem ser interpretados como:

*Q1* e *Q2*: indicam ponto de associação social positiva, no sentido que uma localização possui vizinhos com valores semelhantes.

*Q3* e *Q4*: indicam pontos de associação espacial negativa, no sentido que uma localização possui vizinhos com valores distintos.

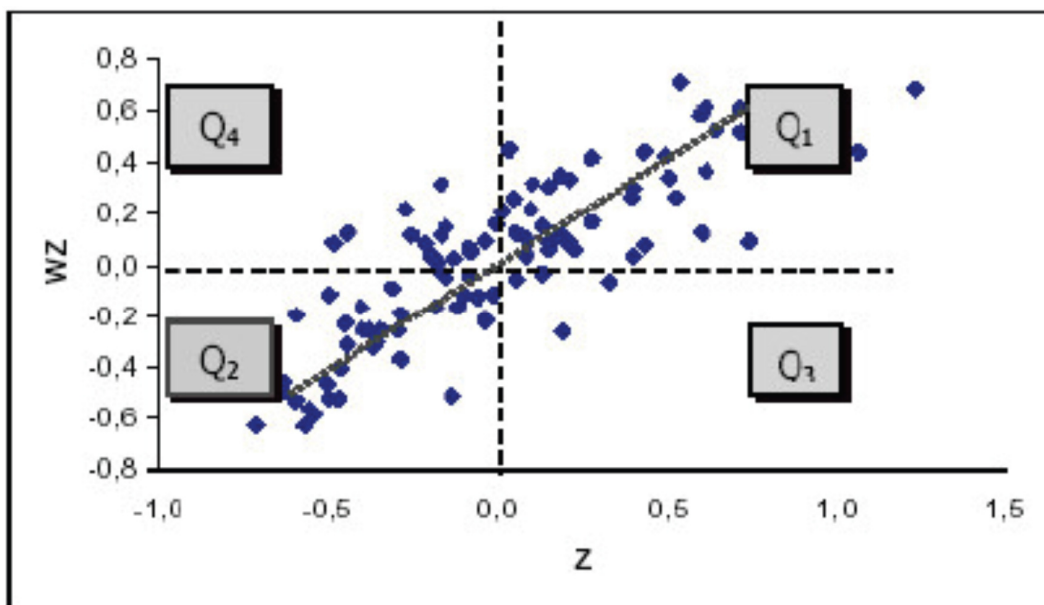


Figura 13. Diagrama de escalonamento de Moran para o índice de inclusão/exclusão social de São Paulo, censo de 1991.

Nota-se que o diagrama de espalhamento de Moran confirma os resultados apresentados, onde indicamos que o índice global de Moran para o indicador de inclusão/exclusão social para os distritos de São Paulo apresentava valor estatisticamente significativo. Como mostrado na Figura 13, a maior parte dos distritos de São Paulo está localizado nos quadrantes *Q1* e *Q2*, que apresentam associação espacial positiva. Os pontos localizados nos quadrantes *Q3* e *Q4* podem ser vistos como regiões que não seguem o mesmo processo de dependência espacial nas duas escalas de análise.

O Índice de Moran  $I$  é equivalente ao coeficiente de regressão linear que indica a inclinação da reta de regressão de  $wz$  em  $z$ . O diagrama de



espalhamento de Moran também pode ser apresentado na forma de um mapa temático bidimensional, no qual cada polígono é apresentado indicando-se seu quadrante no diagrama de espalhamento como ilustra a Figura 14, em que mostramos o mapa de espalhamento do Índice de Moran para o índice de inclusão/exclusão social da cidade de São Paulo (Sposati, 1996). Nesta figura, “Alto-Alto”, “Baixo-Baixo”, “Alto-Baixo” e “Baixo-Alto” indicam, respectivamente, os quadrantes  $Q1$ ,  $Q2$ ,  $Q3$  e  $Q4$ , mostrados na Figura 13. Mostra-se uma forte polarização centro-periferia e observa-se que os distritos localizados nos quadrantes  $Q3$  e  $Q4$  (indicados pela cor azul) podem ser entendidos como regiões de transição entre o centro da cidade (que tende a apresentar valores positivos do índice de inclusão/exclusão social) e as duas grandes periferias de São Paulo (zona Sul e zona Leste).

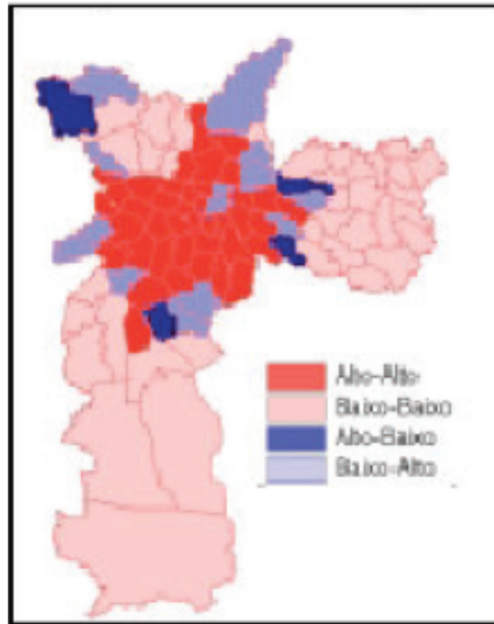


Figura 14. Mapa de espalhamento de Moran para o índice de inclusão/exclusão social da cidade de São Paulo, censo 1991.

- Indicadores locais de associação espacial. Estes índices produzem um valor específico para cada área, permitindo assim, a identificação de agrupamentos. O índice local de Moran pode ser expresso para cada área  $i$  a partir dos valores normalizados  $z_i$  do atributo:

$$I_i = \frac{z_i \sum_{j=1}^n w_{ij} z_j}{\sum_{j=1}^n z_j^2} \quad (60)$$

A significância estatística do uso do índice de Moran local é calculada de forma semelhante ao índice global. Para cada área, calcula-se o índice local, e depois permuta-se aleatoriamente o valor das demais áreas até obter uma pseudo distribuição para a qual podemos calcular os parâmetros de significância. Uma vez determinada a significância estatística do índice local de Moran, é interessante gerar um mapa indicando as regiões que apresentam correlação local significativamente diferente do resto dos dados. Estas regiões podem ser vistas como “bolsões” de não estacionariedade, pois, são áreas com dinâmica espacial própria e que merecem análise detalhada (Anselin, 1995; Anselin, 1996). Para o uso do índice de exclusão/inclusão social da cidade de São Paulo (censo de 1991). Esse mapa da Figura 15 mostra claramente os agregados de pobreza e de situação crítica, onde o agravamento das condições sociais resulta numa degradação significativa das condições de vida.

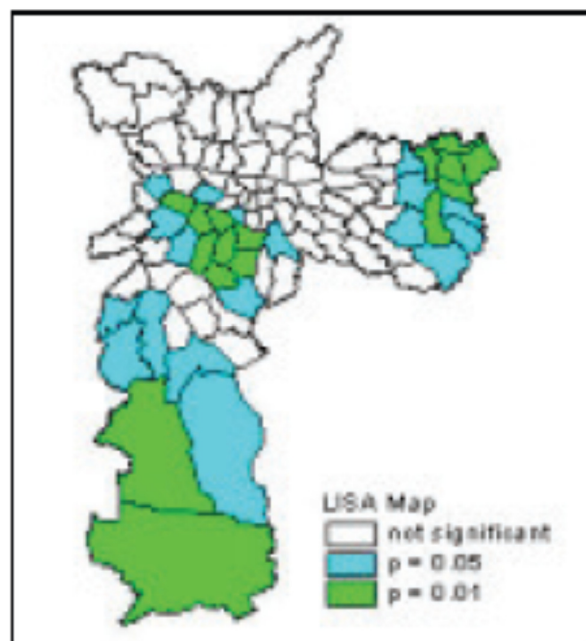


Figura 15. Indicador de auto correlação parcial para o índice de exclusão/inclusão da cidade de São Paulo (censo de 1991), para os valores com significância maior que 95%.

Para este projeto pretendemos propor e executar diferentes técnicas de análise considerando aplicativos como R nos seus módulos GeoR e Geoglm, Terraview e Spring para as variáveis constituídas por diferentes deficiências e variáveis relacionadas pertencentes a blocos como: instrução, identificação, trabalho, família, condições de moradia e posse de outros bens; bem como; estudos comparativos entre variáveis

pertencentes a diferentes deficiências e variáveis aos diferentes blocos mencionados anteriormente.

d) Estimação de indicadores

Um dos problemas básicos em dados agregados por área é que, para uma mesma população estudada, a definição espacial das fronteiras das áreas afeta os resultados obtidos. As estimativas obtidas dentro de um sistema de unidades de área são função das diversas maneiras que essas unidades possam ser agrupadas; pode-se obter resultados diferentes simplesmente alterando as fronteiras destas zonas destas zonas. Este problema é conhecido como “problema da unidade de área modificável”.

Em muitos dos estudos envolvendo dados de área, o dado agregado é a única fonte disponível, porém o objeto de estudo diz respeito a características e relacionamentos individuais. Alguns destes estudos procuram estabelecer relações de causa-efeito entre diferentes medidas, como o uso de modelos de regressão; um exemplo clássico é correlacionar anos de estudo de estudo do chefe de família e sua renda, que usualmente apresenta forte correlação. Nota-se, no entanto que devido aos efeitos de escala e de agregação de áreas, os coeficientes de correlação podem ser inteiramente diferentes no indivíduo e nas áreas. Este fenômeno nas ciências sociais e na epidemiologia é chamada de “falácia ecológica”.

Para que possa contornar o problema da “falácia ecológica” (Openshaw, 1997), recomenda-se utilizar a melhor resolução espacial disponível juntamente com um tratamento adicional dos dados, principalmente nos casos de pequenas áreas em que calculamos taxa sobre um universo populacional reduzido (Openshaw and Avandis, 1999).

Este tipo de problema é típico de recobrimentos espaciais sobre divisões políticas administrativas, onde se analisam áreas com valores muito distintos de população em risco. Vários estudos tem mostrados que em divisões políticas como bairros e municípios apresentam relações inversas de área e população, isto é, os maiores bairros em população tendem a ter menores áreas, e vice-versa. Por isso mesmo, frequentemente o que mais chama atenção em um mapa temático de taxas, são os valores extremos, muitas vezes resultados de um número reduzido de observações, sendo portanto, menos confiável, isto é, apenas flutuação aleatória, conforme exemplo contido na Figura 16 que apresenta um mapa temático com a mortalidade infantil da cidade de Rio de Janeiro, em 1994. Neste mapa, o Rio está dividido em 148 bairros, e a taxa de mortalidade infantil anual para cada bairro, expressa o número de óbitos de menores de 1 ano, por mil nascidos vivos (Cruz, 1996; D’orsi and Carvalho, 1998).

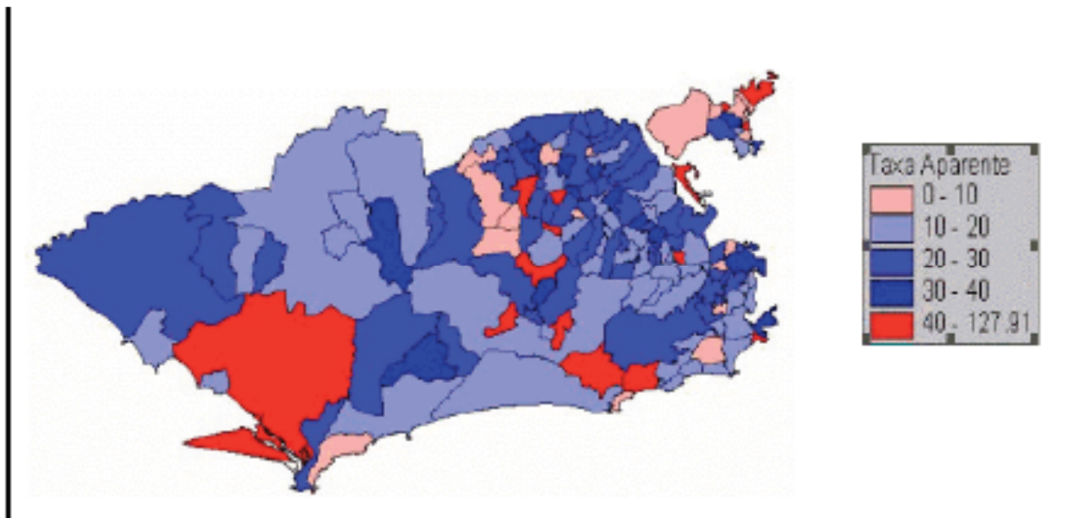


Figura 16. Taxa total de mortalidade infantil por mil nascidos vivos no Rio de Janeiro em 1994.

Tal problema é típico de recobrimentos espaciais sobre divisões político-administrativas, onde se analisam áreas com valores muito distintos de população em risco. Vários estudos tem mostrado que em divisões políticas como bairros e municípios apresentam relações inversas de área e população, isto é, os maiores bairros em população tendem a ter menores áreas e vice-versa. Por este motivo, frequentemente o que mais chama a atenção num mapa temático de taxas, que são os valores extremos, muitas vezes resultados de um número reduzido de observações, sendo portanto menos confiável, isto é, apenas flutuação aleatória.

Para que possa suavizar a flutuação aleatória, considera-se que a taxa estimada pela divisão simples entre contagens de óbitos e de população, taxa observada, é apenas uma realização de um processo não observado, e que é tanto menos confiável quanto menor a população. Assim, propõem reestimar uma taxa mais próxima do risco real ao qual a população está exposta.

É razoável supor que as taxas das diferentes regiões estão auto correlacionadas, e levar em conta o comportamento dos vizinhos para estimar uma taxa mais realista para as regiões de menor população. Esta formulação sugere o uso de técnicas de estimação bayesiana (Gelman et al., 1995). Neste contexto, considera-se que a taxa “real”  $\theta_i$  associada a cada área não é conhecida, e dispomos de uma taxa observada  $t_i = z_i/n_i$ , onde  $n$  é o número de pessoas observadas, e  $z_i$  é o número de eventos na  $i$ -ésima área.

A ideia do estimador bayesiano é supor que a taxa  $\theta$  é uma variável aleatória, que possui uma média  $\mu_i$  e uma variância  $\sigma_i^2$ . Pode ser demonstrado que o melhor estimador bayesiano é dado por uma combinação linear entre a taxa observada e a média  $\mu_i$ :

$$\hat{\theta}_i = w_i t_i + (1 - w_i) \mu_i \quad (61)$$

O fator  $w_i$  é dado por:

$$w_i = \frac{\sigma_i^2}{\sigma_i^2 + \mu_i/n_i} \quad (62)$$

O peso  $w_i$  é tanto menor quanto menor for a população em estudo da  $i$ -ésima área e reflete o grau de confiança a respeito de cada taxa. Para o caso de populações reduzidas, a confiança na área observada diminui e a estimativa da taxa se aproxima de nosso modelo a priori (isto é, se aproxima de  $\mu$ ). Regiões com populações mais baixas terão uma correção maior, e regiões populosas terão pouca alteração em sua taxa. Logo,  $\theta_i$  será estimado, quando  $n$  for pequeno, com maior peso da média da vizinhança.

Neste ponto, deve-se observar que a formulação bayesiana requer as médias e variâncias  $\mu_i$  e  $\sigma_i^2$  para cada uma das áreas. A abordagem mais simples para tratar a estimação desses parâmetros é o estimador bayesiano empírico (Marshall, 1991; Assunção, 2001; Bailey, 2001). Este estimador parte da hipótese que a distribuição da variável aleatória  $\theta$  é a mesma para todas as áreas; isto implica que todas as médias e variâncias são iguais. Pode-se então estimar  $\mu_i$  e  $\sigma_i^2$  diretamente a partir dos dados. Neste caso, calcula-se  $\mu_i$  a partir das taxas observadas:

$$\hat{\mu} = \frac{\sum y_i}{\sum n_i} \quad (63)$$

E estima-se a variância  $\sigma_i^2$  a partir da variância das taxas observadas com relação à média estimada:

$$\sigma^2 = \frac{\sum n_i (t_i - \hat{\mu})^2}{\sum n_i} - \frac{\hat{\mu}}{n} \quad (64)$$

As regiões terão suas taxas ré estimadas aplicando-se uma média ponderada entre o valor medido e a taxa média global, em que o peso da média será inversamente proporcional à população da região.

O estimador bayesiano empírico pode ser generalizado para incluir efeitos espaciais. Neste caso, a ideia é fazer a estimativa bayesiana localmente, convergindo em direção a uma média local e não a uma média global. Basta aplicar o método anterior em cada área considerando como “região” a sua vizinhança. Isto é equivalente a supor que as taxas da vizinhança da área  $i$  possuem média  $\mu_i$  e variância  $\sigma_i^2$  comuns. Neste caso, pode-se falar em estimativa bayesiana empírico local.

Na Figura 17, apresenta-se a detecção da hanseníase em Recife onde foi utilizado um método local para estimar a taxa da doença nos bairros da cidade. Através do mapa corrigido foi possível indicar bairros prioritários para a atuação da vigilância epidemiológica por apresentarem valores altos mesmo após suavização do indicador.

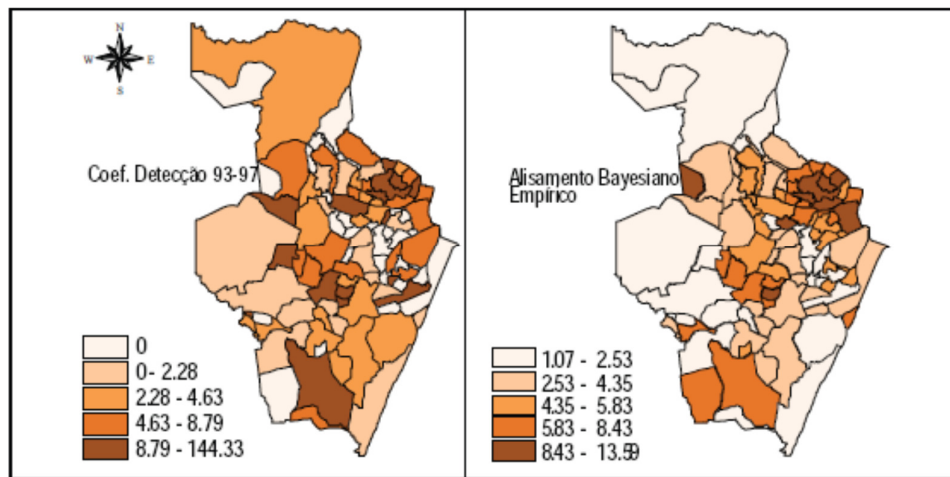


Figura 17. Taxas de detecção média de hanseníases em menores de 15 anos, período 1993-1997, por bairro do Recife, e taxas estimadas através de alisamento bayesiano.

Para este estudo pretendemos calcular e utilizar estimadores bayesianos empíricos para as diferentes deficiências e para as variáveis pertencentes aos diferentes locos.

e) Modelos de regressão

Para este trabalho, vamos considerar os diversos tipos de modelo de regressão que permite incorporar os efeitos espaciais.

- Modelos com efeitos especiais globais

Trata-se de uma classe de modelos de regressão espaciais mais simples, que supõe que é possível capturar a estrutura de correlação espacial num único parâmetro, que é

adicionado ao modelo de regressão tradicional. Neste caso, tem-se duas alternativas para tratar a auto correlação global em um modelo de regressão. Na primeira, a auto correlação ignorada é atribuída à variável dependente  $Y$ . Esta abordagem é denominada como modelo espacial auto regressivo misto, dado que se considera a dependência espacial através da adição ao modelo de regressão de um novo termo na forma de uma relação espacial para a variável dependente, Formalmente isto é expresso como:

$$Y = \rho WY + X\beta + \varepsilon \quad (65)$$

onde  $W$  é a matriz de proximidade espacial, e o produto  $WY$  expressa a dependência espacial em  $Y$  e  $\beta$  é o coeficiente espacial auto regressivo. A hipótese nula para a não existência de auto correlação, é que  $\rho = 0$ . A ideia básica neste modelo é incorporar a auto correlação espacial como componente do modelo. Em termos de componentes individuais, este modelo pode ser expresso como

$$y_i = \rho \left( \sum_j w_{ij} y_j \right) + \sum_{i=1} x_i \beta_i + \varepsilon_i \quad (66)$$

O segundo tipo de modelo de regressão espacial com parâmetros globais considera que os efeitos espaciais são um ruído, ou perturbação, ou seja, fator que precisa ser removido. Neste caso, os efeitos de auto correlação espacial são associados ao termo de erro  $\varepsilon$  e o modelo pode ser expresso por:

$$Y = X\beta + \varepsilon, \quad \varepsilon = \lambda W + \xi \quad (67)$$

onde  $W\varepsilon$  é a componente do erro com efeitos espaciais,  $\lambda$  é o coeficiente auto regressivo e  $\xi$  é a componente do erro com variância constante e não correlacionada. A hipótese nula para a não existência de auto correlação é  $\lambda = 0$ , ou seja, o termo do erro não é espacialmente correlacionado.

Na prática, a distinção entre os dois tipos de modelos de regressão espacial com parâmetros globais é difícil, pois, apesar de diferenças na sua motivação, eles são muito próximos em termos formais. Estes modelos estão incluídos em ambientes de estatística espacial avançados como *SpacerSat* (Anselin, 1992), *S-Plus* e *R*, esse último de domínio público.

Os modelos de regressão espacial com efeitos globais partem do princípio de que o processo espacial subjacente aos dados analisados é estacionário. Isto implica

que os padrões de auto correlação espacial existentes nos dados podem ser capturados num único parâmetro.

- Diagnóstico de modelos com efeitos espaciais

A análise gráfica do resíduo é o primeiro passo para avaliar a qualidade do ajuste da regressão, tratando-se de uma etapa importante no diagnóstico do modelo, buscando indícios de rupturas de pressupostos de independência. Uma alta concentração de resíduos positivos (ou negativos numa parte do mapa é um bom indicador da presença de auto correlação espacial. Para um teste quantitativo, o mais comum é utilizar o índice  $I$  de Moran.

Como os estimadores e os diagnósticos tradicionais de regressão não levam em conta os efeitos espaciais, as inferências, como por exemplo, as indicações de qualidade de ajuste baseada em coeficiente de determinação, serão incorretas. Estas consequências são similares às que acontecem quando uma variável explicativa significativa é omitida do modelo de regressão. Quando se quer comparar um ajuste obtido por um modelo de regressão padrão, com um ajuste obtido por um dos modelos cuja especificação considera a auto correlação espacial, uma medida como o coeficiente de determinação não é mais confiável.

O método mais usual de seleção de modelos de regressão baseia-se nos valores de máxima verossimilhança dos diferentes modelos, ponderado pela diferença no número de parâmetros estimados. Nos modelos com estrutura de dependência espacial ou temporal utiliza-se os critérios de informação onde a avaliação do ajuste é penalizada por uma função do número de parâmetros. Cabe ainda observar que é necessário ainda levar em conta o número de parâmetros independentes ao se incluir funções espaciais nos modelos. Para cada nova variável no modelo de regressão, acrescenta-se um parâmetro.

$$AIC = -2LIK + 2k \quad (68)$$

onde  $LIK$  é o logaritmo da verossimilhança maximizado e  $k$  é o número de coeficientes de regressão. Segundo este critério, o melhor modelo é o que possui menor valor de  $AIC$ . Diversos outros critérios de informação estão disponíveis, a maior parte das quais são variações do  $AIC$ , com mudança na forma de penalização de parâmetros ou observações.



Neste projeto pretendemos, também, propor modelagem de regressão simples, múltipla e multivariada considerando como dependente as diferentes deficiência e como independentes as demais.

#### **14 - Risco deficiência**

Segundo a OMS, pessoas com deficiência apresentam altas estimativas de prevalência; números crescentes devido ao envelhecimento da população e por conta do aumento global das condições crônicas de saúde associadas à deficiência tais como diabetes, doenças cardiovasculares e doenças mentais; experiências diversas em que a experiência da deficiência resultante de interação entre condições de saúde, fatores pessoais e ambientais variando largamente; e, por fim, populações vulneráveis cuja a prevalência varia com as condições de cada país, poder aquisitivo das pessoas, condições de trabalho e nível de formação. Fatores como esses são considerados como riscos para que as pessoas sejam portadoras de deficiência, que por sua vez pode agravar as situações citadas anteriormente.

A necessidade de se medir o bem estar surgiu na metade do século XX como consequência da crescente preocupação em termos da qualidade de vida que decorria das consequências do processo de industrialização. No entanto, à medida que a investigação foi avançando nesta área, tem surgido novas razões que justificam a necessidade de avaliar o bem estar ou qualidade de vida, entre os quais considerar a criação de um índice de risco pessoas com deficiência para melhor avaliar a qualidade de vida destas pessoas.

Indicadores de qualidade de vida constituem informações condensadas, simplificadas, qualificadas, que facilitam a comunicação, comparações e o processo de decisão. Os indicadores de qualidade de vida propõem-se ainda, a ser um incentivo para a mobilização da sociedade afim de pressionar os que tomam decisões.

Uma questão de interesse nos dias de hoje é avaliar a qualidade de vida, seja de um indivíduo, município, estado, região ou país e um dos pontos que passaram a compor a avaliação desta qualidade de vida é formado pelo conjunto de variáveis como nível de instrução, tempo de transporte, renda, idade, sexo e outros que podem dependendo do seu valor, aumentar ou diminuir o risco deficiência que pode servir de instrumento de avaliação e na execução de políticas públicas dirigida de forma preventiva a todo o conjunto da população e também para direcionar o trabalho para que o conjunto de pessoas com deficiência seja melhor incluída na sociedade e nos seus diferentes deveres e direitos.

Segundo Meadows (1998), a seleção e utilização de indicadores é um processo cheio de problemas das quais destacam-se: índices muito compostos; medir o que é mensurável em vez de medir o que é importante; dependência de um falso modelo; falsificação deliberada; confiança exagerada, e, por fim; serem incompletos.

Apesar dos índices serem difíceis de definir, não significa que não os deveremos usar, mas, continuar a pesquisar índices mais completos e evolucionários. Para que possam ser melhor utilizados.

Diante deste cenário, tem surgido razões que justifiquem a necessidade de avaliar o bem estar ou qualidade de vida das pessoas com deficiência, propomos a criação do *índice risco pessoa com deficiência* composto pela ponderação das respostas das diferentes variáveis obtidas a partir dos microdados do Censo do IBGE no caso brasileiro e selecionadas como significativas após aplicar para cada uma dos diferentes tipos de deficiência estudada.

Neste projeto, estudaremos e proporemos a criação deste índice que será composto pela ponderação das respostas das diferentes variáveis obtidas a partir dos microdados do Censo de 2010 do IBGE. Essa metodologia surge gradualmente indo das técnicas estatísticas mais simples a técnicas mais elaboradas como as multivariadas como é o caso da Modelagem de Equações Estruturais utilizando scores fatoriais, regressão logística binária, ordinal, estereótipo.

Como alternativas consideraremos a utilização de outros conjuntos de dados e de outras técnicas como:

#### **14.1 - Regressão Logística Binária**

Com o interesse em prever se incide ou não uma determinada deficiência em função de variáveis independentes associadas a identificação, moradia, família, nível de instrução e trabalho, uma opção é aplicarmos regressão logística, caracterizado por uma variável dependente binária (neste caso, se é ou não portador de uma determinada deficiência ou de pelo menos um dos problemas de deficiência estudados).

Neste caso, pretendemos ajustar um único modelo com deficiência como variável resposta (apresentam os valores 0 se não é portador de deficiência e 1, caso contrário) e as variáveis associadas à identificação, moradia, família, educação e trabalho como independentes no modelo.

A regressão logística é baseada na função logística. Temos:

$$p(x) = [\exp(\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n)]^{-1}$$

em que  $p(x)$  representa a probabilidade da resposta ser 1, isto é, do indivíduo a ser entrevistado ser portador de deficiência, e, por fim; os exponenciais dos coeficientes,  $\exp(\beta_i)$ , representam a razão de chances do indivíduo ser portador de deficiência quando aumentarmos a variável  $X_i$  em uma unidade, e mantendo as outras variáveis constantes. Ou seja, a chance estimada do indivíduo ser portador de deficiência é multiplicada por  $\exp(\beta_i)$  a cada unidade de aumento na variável  $X_i$ .

Apesar da variável deficiência ser binária, os valores ajustados pelo modelo podem assumir qualquer valor entre 0 e 1. Esse valor pode ser interpretado como a probabilidade de ser portador de deficiência (ser 1), dadas as variáveis independentes do modelo. Assim, se o valor dessa probabilidade for maior que 0,5, assumimos que esse indivíduo é portador de deficiência.

Sejam  $\mathbf{x} = (x_1, x_2, \dots, x_k)'$  o vetor de valores de atributos que caracterizam um indivíduo e  $\pi(x)$  a proporção de portadores de deficiência em função do perfil dos indivíduos caracterizados por  $x$ . Neste caso o modelo logístico é adequado para definir uma relação entre a probabilidade de um indivíduo ser portador de deficiência e um conjunto de fatores e atributos que o caracterizam. Esta relação é definida pela função ou transformação logito dada pela expressão

$$\frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k,$$

em que  $\pi(x)$  é definida como

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)},$$

e pode ser interpretado como a probabilidade de um indivíduo ser portador de uma determinada deficiência dado as características que possui representadas por  $x$ . No caso da atribuição da categoria não ser portador da deficiência, as interpretações são análogas.

### 14.1.1. - Interpretação dos coeficientes do modelo.

No caso de modelo de regressão logística, é fundamental o conhecimento do impacto causado por cada variável na determinação da probabilidade do evento de interesse.

Uma medida presente na metodologia de regressão logística e útil na interpretação dos coeficientes do modelo, é o odds, que para uma covariável  $x$  é definida como  $\left[ \frac{\pi(x)}{1-\pi(x)} \right]$ . Aplicando a função log no odds tem-se a transformação logito. Para uma variável dicotômica assumindo valores ( $x = 1$ ) e ( $x = 0$ ), obtém-se que o odds é dado por  $\left[ \frac{\pi(1)}{1-\pi(1)} \right]$  e  $\left[ \frac{\pi(0)}{1-\pi(0)} \right]$ , respectivamente. A razão entre os odds em ( $x = 1$ ) e ( $x = 0$ ) define o odds ratio dado por:

$$\psi = \frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))} = \frac{\left( \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}} \right) / \left( 1 - \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}} \right)}{\left( \frac{e^{\beta_0}}{1 + e^{\beta_0}} \right) / \left( 1 - \frac{e^{\beta_0}}{1 + e^{\beta_0}} \right)} = e^{\beta_1}$$

O odds ratio, neste caso, pode ser interpretado como a propensão do indivíduo ser portador de deficiência quando  $x = 1$ , comparado com  $x = 0$  não é portador de deficiência. Por exemplo, seja  $y$  a incidência da deficiência e  $x$  variável indicadora que denota se o indivíduo é alfabetizado ( $x = 0$ ) ou não é alfabetizado ( $x = 1$ ). Se  $\psi = 2$ , podemos afirmar que a incidência da deficiência é duas vezes mais provável nos indivíduos não alfabetizados.

Pretendemos para este projeto, verificar quais as variáveis independentes (exemplo, nível de instrução e outras) melhor se ajustam às variáveis dependentes como problemas para enxergar, locomover, mental e ouvir

### 14.1.2. Regressão Logística Ordinal

Muitas das variáveis de estudo nas ciências sociais e humanas são ordinais. Com frequência, a variável dependente toma valores discretos, ou categorias ordenáveis mas cuja a distância entre elas não é conhecida, nem tão pouco constante. Por exemplo, em estudos epidemiológicos e de severidade de deficiência para ver, ouvir ou movimentar conforme estabelecido no Questionário da amostra no Censo Demográfico de 2010 do IBGE que pode ser classificada em não consegue de modo algum; consegue, mas com muita dificuldade; consegue, mas com um pouco de dificuldade; e, por fim, não apresenta problemas para ouvir, ver ou se movimentar. No caso de deficiência intelectual é dividida em tem ou não tem.

Entre os possíveis modelos de regressão logística ordinal é possível destacar: modelo de chances proporcionais que é mais apropriado e interpretação mais fácil, quando a variável resposta a ser considerada é uma variável contínua que foi categorizada; modelo de razão

contínua que é utilizado em casos em que existe interesse específico em uma categoria de variável resposta; modelo de chances proporcionais parciais que permite que algumas covariáveis possam ser modeladas com a suposição de chance proporcional, e para as outras variáveis em que este pressuposto não seja satisfeito, são incluídos no modelo parâmetros específicos que variam para as diversas categorias comparadas, e, trata-se de uma extensão do modelo de chances proporcionais; e, por fim; modelo estereótipo, proposto por Anderson (1984) que é utilizado em situações em que a variável resposta é uma variável ordinal que não se trata de uma versão discreta de alguma variável contínua.

Para este trabalho, temos como variável resposta: deficiências, visual, ouvir, locomover e intelectual que tratam-se de variáveis ordinais que não são versões de variáveis contínuas, em vista disso, adotamos neste trabalho o modelo estereótipo.

#### 14.1.3. Especificação do modelo estereótipo

Suponhamos que a variável dependente é constituída por  $J$  categorias ( $m = 1, \dots, J$ ) e consideremos  $K$  regressores ( $j = 1, \dots, K$ ). O modelo estereótipo define-se numa fase inicial com o modelo de regressão multinomial ao qual se adiciona a condição  $\beta_{m|j} \equiv \phi_m \tilde{\beta}_j$ , onde  $J$  é a categoria de referência, isto é, temos que o modelo de regressão multinomial é dado por

$$\text{Prob}(y = m|x) = \frac{\exp(\beta'_{m|J} x)}{\sum_{j=1}^J \exp(\beta'_{m|j} x)}, \text{ com } m = 1, \dots, J. \quad (69)$$

Substituindo  $\beta_{m|j} = \phi_m \tilde{\beta}_j$  na equação (55) resulta no modelo estereótipo que pode ser escrito matematicamente por:

$$\text{Prob}(y = m|x) = \frac{\exp(\phi_m \tilde{\beta}' x)}{\sum_{j=1}^J \exp(\phi_j \tilde{\beta}' x)} = \frac{\exp(\phi_m \tilde{\beta}_0 + \phi_m \tilde{\beta}_1 x_1 + \dots + \phi_m \tilde{\beta}_k x_k)}{\sum_{j=1}^J \exp(\phi_m \tilde{\beta}_0 + \phi_m \tilde{\beta}_1 x_1 + \dots + \phi_m \tilde{\beta}_k x_k)}, \text{ com } m = 1, \dots, J. \quad (70)$$

Para alguns parâmetros da equação (56) que não são identificáveis, consideramos como restrições  $\phi_m \tilde{\beta}_0 \equiv \theta_m$  ( $m = 1, \dots, J$ ), onde  $\phi_j = 0$  e  $\phi_m = 1$ ; e  $\phi_m \tilde{\beta}_j \equiv -\theta_m \beta_j$  ( $m = 1, \dots, J$  e  $j = 1, \dots, k$ ), onde  $\phi_j = 0$  e  $\phi = 1$ . Deste modo, a partir da equação (70), o modelo estereótipo pode ser escrito da seguinte forma:

$$\text{Prob}(y = m|x) = \frac{\exp(\theta_m - \phi_m \beta' x)}{\sum_{j=1}^J \exp(\phi_j \beta' x)}, \quad (71)$$

com  $m = 1, \dots, J$  e onde  $\theta_j = 0$ ,  $\phi_j = 0$  e  $\phi = 1$ .

#### 14.1.3.1. Interpretação dos coeficientes estimados

Aplicando logaritmo na função (85) para duas quaisquer categorias obtemos:

$$\log \left[ \frac{p(Y=q/x)}{p(Y=r/x)} \right] = (\theta_q - \theta_r) - (\phi_q - \phi_r) \beta' x. \quad (72)$$

Aplicando a função exponencial a função exponencial à equação (58), segue:

$$\Omega_{q/r} = \frac{p(Y=q/x)}{p(Y=r/x)} = \exp \{ (\theta_q - \theta_r) - (\phi_q - \phi_r) \beta' x \} \quad (73)$$

A equação (59) permite avaliar a razão de chances antes e depois de adicionarmos uma unidade à variável  $x_j$ , isto é,

$$\frac{\Omega_{q/r}(x, x_k + 1)}{\Omega_{q/r}(x, x_k)} = \exp \{ (\phi_r - \phi_q) \beta_x \} \quad (74)$$

O valor obtido na expressão (74) pode ser interpretado como adicionando uma unidade na variável  $x_k$ , a razão de chances da categoria  $r$  varia  $\exp \{ (\phi_r - \phi_q) \beta_x \}$  mantendo todas as outras variáveis constantes.

#### 14.1.3.2. Estimação dos coeficientes estimados

O modelo estereótipo é estimado pelo método de estimação de Máxima Verossimilhança que é obtida pelo sistema de equações dadas em (75) a seguir:

$$p_i = \begin{cases} \text{Prob}(y_i = 1 | x_i, \phi, \theta) & \text{se } y_i = 1 \\ \vdots & \\ \text{Prob}(y_i = m | x_i, \phi, \theta) & \text{se } y_i = m \\ \vdots & \\ \text{Prob}(y_i = J | x_i, \phi, \theta) & \text{se } y_i = J \end{cases} \quad (75)$$

onde  $p_i$  é a probabilidade de observar qualquer valor de  $y$ ; e, a  $\text{Prob}(y_i = 1 | x_i, \phi, \theta)$  foi definida na expressão (71). Ao assumirmos que a amostra é independente e identicamente distribuída, a função de verossimilhança é dada pela expressão (62) a seguir:

$$L(\beta, \phi, \theta | y, x) = \prod_{i=1}^N p_i = \prod_{m=1}^J \prod_{y=m} \text{Prob}(y = m | x, \phi, \theta) \quad (76)$$

em que  $\prod_{y=j}$  indica as multiplicações sobre todos os casos onde  $y = m$  ( $m = 1, \dots, J$ ). Aplicando logaritmo a função de verossimilhança obtida em (76) obtemos o logaritmo da função de verossimilhança dada em (77) logo abaixo:

$$\log(L(\beta, \phi, \theta|y, x)) = \sum_{m=1}^J \sum_{y=m} \log[\text{Prob}(y = m|x, \phi, \theta)] \quad (77)$$

Os parâmetros  $\phi$ s e  $\theta$ s da equação (77) foram estimados pelo método de Newton-Raphson.

A razão de chance formada terá uma tendência de crescimento, já que os pesos podem ser construídos pela ordenação. Assim, o efeito das covariáveis na primeira razão de chances é menor que o efeito na segunda e assim sucessivamente.

A determinação destes pesos pode ser decididos a priori, sendo estimados por algum estudo piloto ou conjunto de valores apropriadamente escolhidos.

Por exemplo, situação de número de deficiências que um indivíduo possui e que pode variar entre 0 a 4 deficiências, neste caso seriam seis opções de resposta.

#### 14.1.3.3. Avaliação da qualidade de ajuste

A importância da qualidade de ajuste é justificada pelo fato de que a falta de ajuste pode levar a vies de estimação de efeitos importantes. A avaliação do ajuste pode detectar: covariáveis importantes, interações omitidas, casos em que a função de ligação (logito) não foi apropriada, casos em que a suposição de chance proporcional foi violada.

A qualidade do ajuste dos modelos ordinais pode ser verificada usando os testes de Pearson ou deviance. Esses testes envolvem a criação de uma tabela de contingência na qual as linhas consistem de todas as possíveis configurações das covariáveis do modelo e as colunas são as categorias de resposta ordinal. As contagens esperadas ( $E_{ij}$ ) dessa tabela são expressas

$$\text{pôr } E_{ij} = \sum_{l=1}^{N_L} \hat{p}_{ij}, \text{ em que } N_L \text{ é o número total de indivíduos classificados na linha } l \text{ e } \hat{p}_{ij}$$

representa a probabilidade de um indivíduo na linha  $l$  ter a resposta  $j$  calculada a partir do modelo adotado.

O teste de Pearson para avaliar a adequação do ajuste compara essas contagens esperadas com as observadas pela fórmula

$$\chi^2 = \sum_{l=1}^L \sum_{j=1}^k \frac{(O_{lj} - E_{lj})^2}{E_{lj}} \quad (78)$$

A estatística deviance também compara contagens observadas ( $O_{ij}$ ) e esperadas, mas através da fórmula:

$$D^2 = 2 \sum_{l=1}^L \sum_{j=1}^k O_{lj} \log \frac{O_{lj}}{E_{lj}} \quad (79)$$

Os testes para avaliar a qualidade de ajuste do modelo são baseadas nas aproximações das estatísticas (70) e (71) para a distribuição qui-quadrado com  $(L - 1)(k - 1)p$  graus de liberdade, em que  $L$  e  $k$  são conforme definidos anteriormente e  $p$  é o número de covariáveis do modelo. Diferenças significativas levam a conclusão de falta de ajuste do modelo aos dados estudados.

Como alternativa, utilizaremos o teste de Wald que é dado por

$$W = (\hat{p} - \hat{p}_0)' \hat{V}_p^{-1} (\hat{p} - \hat{p}_0) \quad (80)$$

em que  $\hat{V}_p$  é o estimador consistente da matriz de variância-covariância do estimador  $\hat{p}$  do vetor de proporção  $\hat{p}$ . Um estimador  $\hat{V}_p$  pode ser obtido por método de linearização.

## 15 - Cadastro pessoas com deficiência

Em estudos sobre deficiência, nota-se uma grande falta de dados rigorosos e comparáveis e evidências sobre programas que funcionam, prejudicando o envolvimento e ação. Conhecer os números das pessoas com deficiência e suas circunstâncias pode melhorar os esforços para a remoção das barreiras incapacitantes e oferecer serviços que permitam que as pessoas com deficiência participem.

Um outro ponto que devemos considerar é que não podemos contar, única e exclusivamente com os dados do Censo Demográfico do IBGE, que muito embora, sejam de suma importância e de interesse, o censo demográfico é realizado pelo IBGE de 10 em 10 anos, intervalo este, considerado bastante longo, no que diz respeito às demandas das pessoas com deficiência como saúde, lazer entre outras que necessitam de mais dados e análise, dados estes, que em muitos casos, não faz parte dos questionários para o censo elaborados pelo IBGE, mas que são importantes para serem levados em conta na questão do risco deficiência.



Adotar a CIF como um marco universal para a coleta de dados relativos às deficiências e as metas das políticas públicas de prover e assegurar às pessoas com deficiência um melhor acesso aos direitos humanos como educação, saúde, lazer, trabalho, transporte entre outros; e, também, para que possa melhor determinar a prevalência da deficiência e captar melhor sua extensão.

Elaboração de questionário apropriado considerando tópicos como identificação; trabalho; condições de moradia; posse de outros bens como micro computador, televisão e carro entre outros; família; tipos de deficiência de acordo com os respectivos código CIF; condições de saúde; lazer; condições de infraestrutura do entorno entre outros não necessariamente existentes no questionário completo do IBGE.

Considerar a possibilidade de cruzamento com outros bancos de dados como ligados aos dados do IBGE, Transporte, Energia, Saúde, Educação, Trabalho e Emprego, Cultura entre outros; o que permite uma maior agilidade e diminuição de custos para obtenção de dados mais amplos sobre pessoas com deficiência.

A existência desse cadastro pode permitir de forma mais eficiente e de baixo custo, obter dados mais amplo sobre pessoas com deficiência.

Considerar o preenchimento de um questionário mais amplo e que inclua outras variáveis de interesse como questões ligadas a saúde e lazer que não constam no censo, nem mesmo no questionário completo, bem como, promover levantamentos mais detalhados.

Para a construção dos arquivos de banco de dados e formulários, um possível aplicativo que pode ser utilizado é o Epi-data, que compõem o Epi-info que é de domínio público e distribuído pelo CDC de Atlanta.

## **16 – Formas de análise de resultados**

Ao aplicar análise exploratória nos dados, pretendemos obter as distribuições univariada para cada uma das variáveis a serem analisadas e os relacionamentos para cada par de variáveis com o seu respectivo teste de homogeneidade.

No caso da utilização das técnicas de amostragem teremos tamanho da amostra que seja representativa e o conjunto dos dados de indivíduos a serem utilizados na continuidade da análise.

Para a modelagem de equações estruturais a análise é feita avaliando as medidas de qualidade de ajuste para os diferentes modelos e o relacionamento entre as variáveis mediante inspeção gráfica de cada um dos diferentes modelos.

Para o índice risco deficiência, obtido, disponível e monitorado em tempo real e repetido de forma contínua, de modo que a sua evolução possa ser estudada, facilitando o trabalho que deve ser feito pelas autoridades públicas para que possam oferecer soluções começando pelos problemas tidos como mais graves.

Por fim, aplicaremos a modelagem geoestatística poderemos obter a distribuição espacial de cada um dos diferentes tipos de deficiência e comparar com a infraestrutura de cada local em estudo.

### **17 – Criação do Grupo de Pesquisa Estudos Quantitativos, Qualitativos e Comparativos da Discapacidade.**

Além da proposição e desenvolvimento deste projeto, necessitamos que exista um grupo de pesquisa, que além deste projeto em particular, venha desenvolver outros projetos de pesquisa, treinar e desenvolver outros pesquisadores na área de deficiência com mais conhecimentos em termos quantitativos; fortalecer habilidades em disciplinas como epidemiologia, estudos sobre deficiência, educação especial, economia, sociologia e políticas públicas; e, finalmente; oportunidades internacionais de pesquisa e aprendizado formando uma conexão entre as diferentes universidades.

Criação, implementação e participação em grupos de pesquisas que envolve estudos sobre a incapacidade com o objetivo de proporcionar trocas de ideia e de conhecimento entre os pesquisadores considerando outros conjuntos de dados do Brasil e do exterior, bem como, proposição de outros projetos considerando outras metodologias de análise e que possam atender outros objetivos a serem propostos.

Empregar uma “abordagem para dificuldades funcionais” para que possa melhor determinar a prevalência da deficiência e captar melhor a sua extensão.

A pesquisa é essencial para o aumento da compreensão pública sobre questões relacionadas a deficiência, a oferta de informações para a elaboração de programas e políticas dedicados à deficiência, e para alocação eficientes de recursos.

Aumentar a base de dados sobre deficiência a nível nacional e internacional e como mensurá-los; a qualidade de vida e bem estar das pessoas com deficiência; o que funciona para superação de barreiras nos diferentes contextos; e a efetividade e resultados de serviços e programas para pessoas com deficiência.

Em termos estatísticos, propomos metas como: melhora das estatísticas nacionais sobre deficiência; melhorar a comparabilidade entre os diferentes conjuntos de dados, sejam eles a nível nacional ou internacional; e, finalmente; desenvolver ferramentas apropriadas e preencher lacunas existentes entre diferentes pesquisas.

Pesquisas dedicadas à deficiência podem igualmente reunir informações mais abrangentes sobre características das deficiências, tais como prevalência, condições de saúde associadas à deficiência, uso e necessidade de serviços, qualidade de vida, oportunidades e necessidades de reabilitação.

Propor ao IBGE, melhorias que podem ser feitas para os dados do censo que diz respeito a um melhor aproveitamento e possibilidades de outros tipos de análise que diz respeito as condições das pessoas com deficiência como: ao perguntar se determinado respondente possui deficiência perguntar com que idade foi adquirida e observando que ao responder idade zero interpretar que foi adquirida ao nascer e a sua respectiva CIF. Estes acréscimos é motivado de que segundo especialistas de que quando a deficiência foi adquirida ao nascer tem melhores possibilidades de adaptação de quando é adquirida mais tarde e o CIF como uma forma de padronizar os dados nacionais de acordo com o protocolo internacional estabelecido pela OMS e protocolo de Washington.

#### **18 - Benefícios que podem ser oferecidos por esta proposta**

Os resultados deste projeto pode servir de suporte para um maior avanço na inclusão escolar das pessoas com deficiência; melhorar as condições gerais de acessibilidade nos municípios, nos estados e na união; reafirmar paradigmas que não reforcem mais o estereótipo associado a estas pessoas com a finalidade de avançar o acesso das pessoas com deficiência no mercado de trabalho, principalmente em regiões que pessoas com deficiência apresentam maior número.

O conhecimento das variáveis que mais influem a incidência das diferentes deficiências, pode auxiliar na implementação de políticas públicas para melhor prevenir ou inibir diversos fatores de risco que faz aumentar as diferentes deficiência, contribuindo que estas pessoas ganhem mais em termos de inclusão e acessibilidade.

A adoção do cadastro deficiência pode possibilitar uma maior agilidade e uma melhor possibilidade de outros tipos de análises serem feitas sem ficar dependendo exclusivamente das

coletas de dados fornecidos por outros institutos como o IBGE que costuma ter um intervalo grande de anos entre uma pesquisa e outra.

Avaliação crítica do questionário é interessante no sentido de verificar se outras variáveis deveriam ser incluídas no questionário completo, ou, se seria interessante a existência de um censo em separado das pessoas com deficiência, este tipo de avaliação pode servir de instrumento para que os procedimentos do IBGE possam ser melhorados, o que pode contribuir para um melhor aprimoramento dos questionários e da coleta de dados elaboradas pelos recenseadores do IBGE.

A criação do grupo de pesquisa Grupo de estudos quantitativos, qualitativos e comparativos sobre discapacidade pode possibilitar mais interação, relacionamento entre diferentes pesquisadores, desenvolvimento de outras pesquisas e aumento da divulgação.

Como principal resultado deste projeto de pesquisa pretende-se produzir artigos científicos. Tais artigos serão submetidos a periódicos nacionais e internacionais. Planeja-se também apresentar os resultados das investigações conduzidas no âmbito deste projeto em congressos nacionais e internacionais. Desta forma, a difusão dos resultados obtidos no decorrer da execução deste projeto será feita principalmente através de publicações acadêmicas e de apresentações em eventos científicos.

## **19 - Referências bibliográficas**

AKAIKE, H. A new look at the statistical model identifications. *IEEE Transaction Control*. AC(19):716-723, 1974.

AMEMIYA, T. Quantitative response models: a survey. *Journal of Economic Literature*. 19:483-536, 1981.

ANDERSON, J.; GERBING, D.W. The effects of sampling error on convergence, improper solutions and goodness-of-fit indices maximum likelihood confirmatory factor analysis. *Psychological Bulletin*. **88**:588-1600, 1980.

ANSELIN, L. SpaceStat tutorial: a workbook for using SpaceStat in the analysis of spatial data. Santa Barbara, NCGIA (National Center for Geographic Information and Analysis), 1992.

ANSELIN, L. Local indicators of spatial association - LISA. *Geographical Analysis* **27**:91-115, 1995.

ANSELIN, L. The Moran scatterplot as ESDA tool to assess local instability in spatial association. In: M. Fisher, H. J. Scholten and D. Unwin (ed). *Spatial Analytical Perspectives on GIS*. London, Taylor & Francis, 1996. 111-126.

ARBUCKLE, J.L. *Amos user's guide*. USA. Small Walters Corporation, 1997.

ASSUNÇÃO, R. Estatística Espacial com Aplicações em Epidemiologia, Economia e sociologia. São Carlos, SP, UFScar, 2001. Disponível na homepage [www.est.ufmg.br/~assuncao](http://www.est.ufmg.br/~assuncao).

BAILEY, T. Spatial Statistics Methods in Health. Cadernos de Saúde Pública **17**:5, 2001.

BAILEY, T. and A. GATTREL. Spatial Data Analysis by Example. London, Longman, 1995.

BARROSO, L.P. *Imputação de dados em painéis para populações finitas*. Tese de Doutorado em Estatística, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo SP, 1995.

BENTLER, P.M.; BONETT, D.G. Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychometrika*, 49:155-173, 1984.

BERGAMO, G.C. *Imputação múltipla livre de distribuição a decomposição por valor singular em matriz de interação*. Tese de Doutorado em agronomia, concentração em Experimentação estatística, ESALQ - Universidade de São Paulo, Piracicaba – SP, 2007.

BISHOP, T.; FIENBERG, S.; HOLLAND, P. *Discrete multivariate analysis*. Cambridge, MA:MIT Press, 1975.

BISTAFFA, B.C. *Incorporação de indicadores categóricos ordinais em modelos de equações estruturais*. Dissertação de mestrado. Instituto de Matemática e Estatística, USP, São Paulo-SP, 2010.

BLACK, H.M. Four variables causal models and partial correlations. *American Journal of Sociology*, 168:182-194, 1962.

BLANGERO, J.; WILLIAMS, J.T.; ALMASY, L. Variance component methods for detecting complex trait locos. *Genetic* 42:151-181, 2001.

BOCK, R. D.; BARGMANN, R.G. Analyses of covariance structures. *Psychometrika*, 31:507-534, 1989.

BOLLEN, K.A. *Structural equations with latent variables*. USA: John Wiley, 1989.

BROMAN, A. MATEUS-FIGUERAS, G.; PAWLOWSKY-GLAHN, V. *Identifying quantitative trait locos in experimental cross*. Thesis, University of California, Berkeley, 1997.

BROMAN, K.W. Review of Statistical Methods for QTL Mapping in Experimental Crosses. *Lab. Animal*, 30(7):44-52, 2001.

BROWNE, M.W. *Covariance Structures. Topics in Applied Multivariate Analysis*. D.M. Hawkins (ed.). Cambridge: Cambridge University Press, 1982.

BROWNE, M.W.; CUDECK, R. Alternative ways of assessing model fit. In: Bollen, K. A. & Long, J. S. (Eds.) *Testing Structural Equation Models*. p. 136-162. Beverly Hills, CA: Sage, 1993.

BRUNELLI, R.T. *Análise do impacto de perturbações sobre medidas de qualidade de ajuste para modelos de equações estruturais*. . Dissertação de mestrado. Instituto de Matemática e Estatística, USP, São Paulo-SP, 2012.

BRUNSDON, C. A.S. FOTHERINGHAM AND M.E. CHARLTON, Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, 28(4), 281-298, 1996.

- BUNCH, J.R., NIELSEN, C.P.; SORENSEN, D.C. Rank one modification of the symmetric eigen problem. *Numerische Mathematik*, 31:31-48, 1978.
- BURNHAM, K.P.; ANDERSON, D.R. *Model selection and multimodel inference*. USA: Springer, 1998.
- BURNHAM, K.P.; ANDERSON, D.R. *Model selection and inference*. USA: Springer, 2002.
- BUSSAB, W.; BOLFARINE, H. *Elementos de amostragem*. Profeto Fisher, 2005.
- CAMARINHA FILHO, J.A. *Modelos lineares mistos: estruturas de matrizes de variância e covariâncias e seleção de modelos*. Tese de doutorado, ESALQ, USP, Piracicaba-SP, 2008.
- CANTON, A.W.P. *Análise de Dados Categorizados*. Associação Brasileira de Estatística, 4ª SINAPE, São Paulo-SP, 1980.
- CARLBÖRG, O.; ANDERSON, L.; KINGBORN, B. The use a genetic algorithm for simultaneous mapping of multiple interactions quantitative trait loci. *Genetic*, 155:2003-2010, 2000.
- CARVALHO, J. O. F. *Soluções tecnológicas para viabilizar o acesso do deficiente visual à Educação à Distância no Ensino Superior*. Faculdade de Engenharia Elétrica e Computação, UNICAMP, Campinas-SP, 2001.
- CARVALHO, M.L.; NATÁRIO, L.C.; *Análise de dados espaciais*. Sociedade Portuguesa de Estatística, Lisboa, Portugal, 2010.
- CASELLA, G.; BERKER, P.L., 1990. *Statistical inference*. Wadsworth, Brooks, California, EUA.
- COCHRAN, W.G. *Sampling Techniques*. Third edition, New York, John Wiley & Sons, 1977.
- CRUZ, O. C. Homicídios no Estado do Rio de Janeiro: análise da distribuição espacial e sua evolução. Dissertação de mestrado/Faculdade de saúde Pública-USP, 1996.
- DIGGLE, P. *Spatial statistics in the biomedical science: future directions*. Lancaster, Lancaster University, 2001.
- DIGGLE, P.; RIBEIRO JR., P.J. *Model based geostatistics*. Springer Series in Statistics, London, UK, 2007.
- DINIZ, C.; LOUZADA, F. *Modelagem estatística para risco de crédito*. Associação Brasileira de Estatística, João Pessoa-PB, 2012.
- DOMINGOS, M.A. *Sentidos e significados produzidos por alunos e professores no cotidiano numa escola do sistema regular de ensino a partir da inclusão de alunos portadores de necessidades educacionais especiais*. Dissertação de Mestrado, Faculdade de Educação, PUC-MG, Belo Horizonte-MG, 2006.
- D'ÓRSI, E. and M. S. CARVALHO. Perfil de Nascimentos no Município do Rio de Janeiro - Uma Análise Espacial. *Cadernos de Saúde Pública* v.14, n.1, p.367-379, 1998.
- DRAPER, N.R.; SMITH, H. *Applied regression analysis*. Segunda edição, Nova York, Wiley, 1998.
- EUCLIDES-FILHO, K. *Melhoramento genético animal no Brasil: fundamentos, história e importância*. Embrapa-DF, 2000.
- FIGUEIRA, E. *Caminhando em Silêncio*. Giz Editorial e Livraria Ltda, São Paulo – SP, 2008.

- FISHER, R.A. On the mathematical foundations of theoretical statistics. *Royal Society of London Philosophical Translations (Series A)*, **222**:309-368, 1922.
- FONTOURA, D.S. *Inserção de pessoas com deficiência no mercado de trabalho*. Tese de Doutorado, Escola de Administração, UFRS, Porto Alegre-RS, 2006.
- FOTHERINGHAM, A.S., C. BRUNSDON AND M.E. CHARLTON, 2000, *Quantitative Geography*, London: Sage
- FOTHERINGHAM, A.S., M.E. CHARLTON AND C. BRUNSDON, The Geography of Parameter Space: An Investigation into Spatial Non-Stationarity. *International Journal of Geographic Information Systems*, **10**: 605-627, 1996.
- GARCIA, V.G. *Pessoas com deficiência e o mercado de trabalho – histórico*. Tese de Doutorado, Instituto de Economia – UNICAMP, Campinas – SP, 2010.
- GASPARETTO, M.G.R.F.; TEMPORINI, E.R.; CARVALHO, K.M.M.; KARA-JOSÉ, N. O aluno portador de visão subnormal na escola regular: desafio para o professor? *Arquivo Brasileiro de Oftalmologia*, 64:45-51, 2001.
- GELMAN, A., CARLIN, J.B., STERN, H.S., RUBIN, D.B. (1995) *Bayesian Data Analysis* Chapman & Hall/CRC.
- GILKS, W.R., RICHARDSON, S., SPIEGELHALTER, D.J. (orgs) (1998), *Markov Chain Monte Carlo in Practice*, Chapman & Hall.
- GOLDBERGER, A.S.; DUCAN, O.D. *Structural equation models in the social sciences*. New York: Seminar Press/Harcourt Brace, 1973.
- GOODMAN, A.S. Structural equation modeling in the social science. *Econometrica*, 40:979-1001, 1972.
- GOOLUB, G.H.; REINSCH, C. Singular value decomposition and least squares solutions. *Numerische Mathematik*, **14**, 403-420, 1970.
- GUTIERREZ, G.C. *Estimação das escalas dos construtos capital cultural e capital econômico e análise no efeito escola nos dados do PERU-PISA, 2000*. Dissertação de mestrado, Departamento de Engenharia Elétrica, PUC, Rio de Janeiro-RJ, 2005.
- OMS, T. The statistical implications of a system of simultaneous equations. *Econometrica*, 11:1-12, 1943.
- HAWKING, S.W. (2011), *Relatório mundial sobre a deficiência*. OMS, São Paulo=SP.
- HOSMER, D.W.; LEMESHOW, S. *Applied Logistic regression*. New York, John Wiley & Sons, 2000.
- JONHSON, R.A.; WICHERN, D.W. *Data Applied Multivariate Statistical Analysis*, New Jersey, Prentice-Hall, USA, 2000.
- JEFFREYS, H. The present position in probability theory. *British Journal for Philosophy of Science*, 5:275-289, 1935.
- JONHSON, R.A.; WICHERN, D.W. *Data Applied Multivariate Statistical Analysis*. New Jersey, Prentice-Hall, USA, 2006.

- JÖRESKOG, K.G. Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32(4):443-482, 1967.
- JÖRESKOG, K.G.; SÖRBOM, D. *LISREL VI: Analysis of linear structural relationships by maximum likelihood and least square methods*. Scientific Software, Mooresville, 1986.
- JÖRESKOG, K.G.; SÖRBOM, D. *LISREL 7: A guide to the program and applications*. SPSS Inc., 1989.
- JÖRESKOG, K.G.; SÖRBOM, D. *LISREL 8: A guide to the program and applications*. SPSS Inc., 1997.
- KASS, R.; RAFTERY, A.Q. Bayes factor and model uncertainty. *Journal of American Statistical Association*, 90:773-795, 1995.
- KRZANOWSKI, W.J. Cross-validation in principal component analysis. *Biometrics*, Washington, 43:575—584, 1987.
- KUTNER, M.H.; NACHTSHEIM, C.J.; NETER, J.; LI, W. *Applied Linear Statistical Models*. McGraw-Hill Irwin., fifth edition, New York, USA, 2005.
- KYUNG, M.; GILL, J.; GLOSH, M.; CASELLA, G. Penalized regression standard errors and bayesian lassos. *Bayesian Analysis*, 5(2): 369-412, 2010.
- LAIRD, N.M. Missing data in longitudinal studies. *Statistics in Medicine*, 7:305—351, 1988
- LATIF, S.A. *Modelagem de equações estruturais*. Dissertação de mestrado, Instituto de Matemática e Estatística, USP, São Paulo-SP, 2000.
- LATTIN, J.; GREEN, P.; CARROLL, J. D. *Análise de Dados Multivariados*. Editora CENAGE, 2011.
- LAZARSELD, P.F.; HENRY, N.W. *Latent structure analysis*. Houghton Mifflin, Boston, 1968.
- LITTLE, R. J.A.; RUBIN, D.B. *Statistical analysis with missing data*. 2ed, New York, New Jersey: John Wiley & Sons, 2002.
- MADDALA, G.S. *Limited – dependent and quantitative variables in econometrics*. Cambridge: Cambridge University Press, 1983.
- MAGALHÃES, M. N.; LIMA, A. C. P. *Noções de Estatística*. EDUSP, São Paulo – SP, 2011.
- MAITI, S.S.; MUKHERJEE, B.N. Two new goodness-of-fit indices for covariance matrices with linear structure. *British Journal of Mathematical and Statistical Psychology*, v. 44, p. 153-180, 1991.
- MALLOWS, C. L. Some Comments on Cp. *Technometrics*, 15:661-675, 1973.
- MARSHALL, R. Mapping disease and mortality rates using empirical Bayes estimators. *Applied Statistics* **40**:283-294, 1991.
- MARTIN, D. *Geographic Information Systems: Socioeconomic Applications*. London, Routledge, 1995.
- MARTIN, D. An assessment of surface and zonal models of population. *International Journal of Geographical Information Systems* **10**:973-989, 1996.



- MARTIN, D. Optimizing census geography: the separation of collection and output geographies. *International Journal of Geographical Information Science* **12**:673-685, 1998.
- McDONALD. Multivariate model: noncentrality and goodness of fit. *Psychological Bulletin*, 107:247-255, 1989.
- MEADOWS, D. Indicators and informations systems for sustainable development. The sustaibility Institute, Hartland VT.
- MEDRONHO, R.A.; BLOCH, K.V.; LUIZ, R.R. WERNECK, G.L. *Epidemiologia*. Editora Atheneu, 2ª edição, São Paulo-SP, 2009.
- MELHADO, T.T. *Medida de ajuste de modelo de equações estruturais*. Dissertação de mestrado, Instituto de Matemática e Estatística, USP, São Paulo-SP, 2004.
- MELHADO, T.T. *Influência local para modelos de equações estruturais com distribuição elíptica*. Tese de Doutorado, Instituto de Matemática e Estatística, USP, São Paulo-SP, 2009.
- MILLER, A. *Subset selection in regression*. Chapman and Hall, New York, 1990.
- MULAIK, S.A.; JAMES, L.R.; VANAISTINE, J.; BENNETT, N.; LIND, S.A.; STILWELL, C.D. Evaluation of goodness-of-fit indices for structural equations models. *Psychological Bulletin*, 105:430-445, 1989.
- MUNDSTOCK, G.C. *Amostragem I. Cadernos de matemática e estatística, série B*. Instituto de Matemática, UFRS, Porto Alegre-RS, 2005.
- NAKAMICH, R.; UKA, Y.,; KISHINO, H. Detection of closely linked multiple quantitative trait loci using a genetic algorithm. *Genetics*, 158:463-475, 2001.
- NATARAJAN, J.M.; LIPSITZ, S.R., FITSMAURICE, G.M.; MOORE, C.G.; GONIN, R. Variance estimation in complex survey sampling for generalized linear models. *Applied Statistics*, 57(1):75-87, 2008.
- OLIVEIRA, P.T.M.S. *Estimação e testes de hipótese em calibração comparativa*. Dissertação de mestrado, Instituto de Matemática e Estatística, USP, São Paulo-SP, 2001.
- OLIVEIRA, P.T.M.S. *Aplicação do algoritmo genético no mapeamento de genes epistáticos em cruzamentos controlados*. Tese de doutorado. Instituto de Matemática e Estatística, USP, São Paulo-SP, 2008.
- OLIVEIRA, P.T.M.S., 2013. Pessoas com deficiência: análise dos resultados do Censo 2010 e a sua evolução. In: 58 RBRAS/ 15 SEAGRO, no período entre 22 a 26 de julho de 2013, Campina Grande – PB.
- OLIVEIRA, P.T.M.S.; PAVAN, J.M. *Seleção de modelos genéticos usando om critério de informação Bayesiano*. 50 Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria, 11 Simpósio de Estatística Aplicada a Experimentação Agrônômica, Londrina-PR, UEL, 2005.
- OPENSHAW, S. Developing GIS-relevant zone-based spatial analysis methods. In: P. Longley and M. Batty (ed). *Spatial Analysis: Modelling in a GIS Environment*. New York, John Wiley, 1997, 55-73.

OPENSHAW, S. and S. ALVANIDES. Applying Geocomputation to the analysis of spatial distributions. In: P. A. Longley, Goodchild, M. F., Maguire, D. J. and Rhind, D. W (ed). Geographical Information Systems: Principles, Techniques, Management and Applications. Chichester, Wiley, 1999, p.267-282.

POLETO, F.Z. Análise de dados categorizados com omissão em variáveis explicativas e respostas. Tese de doutorado. Instituto de Matemática e Estatística, USP, São Paulo-SP, 2011.

PAULA, G.A. *Modelos de regressão com apoio computacional*. Instituto de Matemática e Estatística-USP, São Paulo-SP, 2013.

PAULINO, C.D.; TURKMAN, A.A.; MURTEIRA, B.J.F. *Estatística Bayesiana*. Portugal: Fundação Calouste Gulbenkian, 2003.

PEARL, J. *Probabilistic reasoning intelligent system*. Morgan Kaufmann, 1988.

ROSA, E.R. *Um trabalho das pessoas com deficiência e as relações sociais de produção capitalista: Uma análise crítica da política de cotas no Brasil*. Dissertação de Mestrado, Unoeste, Cascavel – PR, 2009.

RUBIN, D.B. Inference and missing data. *Biometrika*, 63:581—592, 1976.

SAKAMOTO, Y.; ISHGURU, M.; KITAMURA, G. *Akaike information criterion statistics*. Japão: KTK, Scientific Publisher, 1986.

SAMEJIMA, F.A. Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, 17, 1969.

SANTOS, T.M. *Inclusão escolar e educação para todos*. Faculdade de educação, UFRS, Porto Alegre-RS, 2010.

SATORRA, A.; BENTLER, P. M. Some robustness properties of goodness of fit statistics in covariance structure analysis, *1986 ASA Proceedings of the Business and Economic Statistics Section - Alexandria*, VA: American Statistical Association, 549-554, 1986.

SATORRA, A.; BENTLER, P. M. Scaling corrections for chisquare statistics in covariance structure analysis. *ASA 1988 Proceedings of the Business and Economic Statistics Section, Alexandria*, VA: American Statistical Association, 308-313, 1988.

SATORRA, A. Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika*, 54:131-151, 1989.

SATORRA, A.; BENTLER, P. M. Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye and C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research*. Thousand Oaks, CA: Sage, 399-419, 1994.

SCHUSTER, J.; CRUZ, C.D. *Estatística genômica aplicada e populações derivadas de cruzamentos controlados*. Editora UFV, Viçosa-MG, 2004.

SCHWARTZ, G. Estimating the dimension of a model. *The Annals of Statistics*, 6:461-464, 1978.

SEARLE, W. *What is overfitting and how am I avoid it?* <http://www.faqs.org/faqs/ai-faq/newal-nets/part3/section-3.html> (acessado em março de 2004).

- SIALAMPAÄ, J.; CONRADER, J. Model choice in gene mapping: what and why. *Trends in Genetics*, 18:301-307, 2000.
- SILVA, O.M. *A Epopéia Ignorada*. CEDAS, São Paulo-SP, 1986.
- SILVA, N.N. *Amostragem probabilística: um curso introdutório*. EDUSP, São Paulo - SP, 1998.
- SNOW, J. *On the mode of communication of cholera*, 2<sup>nd</sup> edition, 8vo, Londres, 1855.
- STEIGER, J.H.; LIND, J.M. *Statistically based tests for the number of common factors*. Annual Meeting of the Psychometric Society, Iowa City, Iowa, 1989.
- TANAKA, J.S.; HUBA, G.J. A fit index for covariance structure models under arbitrary gls estimation. *British Journal of Mathematical and Statistical Psychology*, 38:197-201, 1985.
- TOLEDO, E.P. *Mapeamento de QTLs utilizando as abordagens clássicas e bayesianas*. Dissertação de mestrado, Escola Superior de Agronomia Luiz de Queiroz, USP, Piracicaba-SP, 2006.
- VIEIRA, K.M.; MILACH, F.T.; HUPPES, D. Equações Estruturais aplicadas à satisfação dos alunos: um estudo no curso de ciências contábeis da Universidade Federal de Santa Maria. *R. Cont. Fin.*, 19(48):65-76, 2008.
- VIEIRA, M.D.T., SKINNER, P.J. Estimating models for panel survey data under complex sampling. *Journal of Official Statistics*, 24(3):343-364, 2008.
- VIEIRA, M.D.T. *Analysis of longitudinal survey data, Saarbruchen*. VDM Verkag, 2009.
- WANG, S., *Simulation study on the methods for mapping quantitative trait loci in Ingrid line crosses*. Theis doctor Philosophy in Genetics and Plant Breeding, Zhejiang University Hanzhou, Zhejiang, China, 2000.
- WANG, S.; BASTEN, C.J.; ZENG, Z-B. *Windows QTL Cartographer v. 2.5*. disponível no site <http://statgen.ncsu.edu/qtlcart/WQTLCart.htm> no dia 01/08/2007.
- WRIGHT, S. Correlation and causation, *Journal of Agricultural Research*. 20:557-585, 1921.
- ZENG, Z-B. *QTL Mapping*. 46 Reunião Anual da Rbras, 9 SEAGRO, Piracicaba-SP, 2001.

