

Vamos detalhar cada fase da construção do script, com exemplos, a teoria por trás (de forma simplificada), e como adaptar para a realidade heterogênea dos bairros de Salvador.

O Objetivo Geral:

O script tenta replicar a metodologia do artigo científico anexado. O objetivo final é **estimar um número mais realista de óbitos por COVID-19** do que o registrado oficialmente. Por quê? Porque os registros oficiais podem ter dois problemas principais:

1. **Sub-inscrição:** Nem todos os óbitos que ocorreram foram oficialmente registrados no sistema.
2. **Classificação Incorreta:** Alguns óbitos que *foram* por COVID-19 podem ter sido registrados com causas vagas ou "genéricas" (Códigos Garbage ou Mal Definidas).

O script simula os passos para corrigir esses dois problemas, usando dados fictícios para ilustrar o processo.

Fase 1: Sub-registro (Estimativa de óbitos "perdidos")

- **Objetivo intuitivo:** Tentar adivinhar quantos óbitos aconteceram, mas "escaparam" do registro oficial.
- **Exemplo:** Imagine que você está contando pessoas entrando numa festa. Se a entrada estiver muito cheia ou você se distrair, pode perder a conta de algumas pessoas. O sub-registro é como tentar estimar quantas pessoas entraram sem você ver.
- **Componente Teórico (Simplificado):** O artigo usa um método demográfico chamado **Equação do Balanço Geral do Crescimento (BGC)**, de Brass. Essa técnica (complexa!) compara a estrutura etária da população (quantas pessoas de cada idade) com a estrutura etária dos óbitos registrados. Se essas estruturas não "batem" de uma forma esperada (considerando nascimentos, mortes e migração), o método estima o quão incompleto está o registro de óbitos. O resultado é um **fator de correção (f)**. Se $f = 1.20$, significa que o método estima que o número real de óbitos é 20% maior que o registrado.
- **Como o Script Faz (Simulação com vamos):**
 - O roteiro *não* implementa o BGC real (que é muito complexo e precisa de dados específicos).

- Em vez disso, ele usa o pacote *vamos* para *simular* uma dinâmica populacional básica para cada bairro/sexo. Ele cria matrizes fictícias de sobrevivência e fertilidade.
- `popdemo::eigscal` calcula a **taxa de crescimento populacional (lambda)** dessas populações simuladas. $\lambda > 1$ indica crescimento, $\lambda < 1$ indica declínio.
- **Ligação Conceitual (Simulada):** O roteiro *assumir* (para fins de demonstração) que bairros com dinâmicas populacionais diferentes (λ alto ou baixo) *poderiam* ter níveis diferentes de sub-registro.
- Ele usa o λ simulado para definir uma *faixa* de onde sortear (`runif`) o fator de correção (Fator_Subregistro). Por exemplo, bairros simulados com λ alto recebem um fator potencialmente maior.
- **Replicando em Salvador (Heterogeneidade):**
 - **Dados Reais:** Você precisaria dos dados populacionais *reais* por bairro, idade e sexo (Censos do IBGE) e dos óbitos *reais* registrados no SIM (Sistema de Informações sobre Mortalidade) também por bairro, idade e sexo, para o período de interesse (ex: 2020, 2021).
 - **Método Real:** Aplicar o BGC (ou outras técnicas de estimativa de sub-registro como métodos baseados em GGD - Generalized Growth Balance) exige conhecimento demográfico. Existem pacotes R que podem auxiliar (pesquisar por "Growth Balance Equation R package"), mas a aplicação correta não é trivial.
 - **Heterogeneidade:** A grande diferença entre bairros de Salvador (renda, acesso à saúde, infraestrutura) significa que o sub-registro *provavelmente varia muito*. Bairros mais pobres ou com piores serviços de saúde podem ter sub-registro maior. Você **não pode** usar um fator único para a cidade toda. Seria necessário calcular (ou estimar com base em características socioeconômicas) um fator (*f*) *específico* para cada bairro ou grupos de bairros similares.
- **Literatura relevante:**
 - **Brass, W. (1975).** *Métodos para estimar fertilidade e mortalidade a partir de dados limitados e defeituosos*. (O clássico sobre BGC).
 - **Hill, K. e outros.** (Vários trabalhos sobre métodos de estimação de mortalidade adulta em países em desenvolvimento).

- **Nações Unidas. (1983).** *Manual X: Técnicas indiretas para estimativa demográfica.* (Referência padrão para técnicas demográficas).
- **Paes, NA (2018).** *Demografia estatística dos eventos vitais.* (Citado no artigo, focado no Brasil).

Fase 2: Redistribuição de Códigos Garbage (CG)

- **Objetivo intuitivo:** Pegar os óbitos registrados como "problema respiratório não especificado" ou "parada cardíaca" (causas muito gerais) durante a pandemia e tentar adivinhar quantos deles foram, na verdade, causados pela COVID-19.
- **Exemplo:** Na sua festa, alguém anota "mal-estar" como motivo de saída de várias pessoas. Se você souber que havia uma comida estragada circulando, pode reclassificar alguns desses "mal-estares" como "intoxicação alimentar". Códigos Garbage são como o "mal-estar" - não dizem a causa raiz.
- **Componente Teórico (Simplificado):** A Classificação Internacional de Doenças (CID-10) tem códigos que são úteis para o sistema de saúde, mas não indicam a **causa básica** da morte (a doença ou lesão que iniciou a cadeia de eventos que levou à morte). Exemplos: Insuficiência respiratória (J96), Sepse (A41. 9), Parada cardíaca (I46). Durante a pandemia, muitas mortes por COVID-19 podem ter sido registradas erroneamente com esses códigos. A metodologia do artigo envolve:
 1. Identificar quais CGs são relevantes no contexto da COVID-19 (baseado em conhecimento médico/epidemiológico).
 2. Decidir como redistribuir uma *parte* desses óbitos CG para a COVID-19 (e outras causas, se aplicável). Existem algoritmos para isso (como os usados pelo Global Burden of Disease - GBD).
- **Como o Script Faz (Simulação):**
 - Filtra óbitos fictícios marcados como "Codigos Garbage".
 - Calcula uma *quantidade* a ser redistribuída (CG_redistribuído_para_COVID) multiplicando o total de CGs observados por uma *proporção aleatória*(Proporcao_Redistribuir_CG_para_COVID). Essa proporção é totalmente simulada.

- **Replicando em Salvador (Heterogeneidade):**
 - **Dados Reais:** Precisa dos dados do SIM com os códigos CID-10 completos para cada óbito.
 - **Método Real:**
 - **Defina uma lista de CG:** Usar listas padronizadas (ex: do GBD, estudos brasileiros como os de França et al.) e adaptá-las ao contexto da COVID-19 na Bahia (talvez com consulta a especialistas locais).
 - **Definir o Método de Redistribuição:** Isso é complexo. Pode ser proporcional aos óbitos já confirmados por COVID em cada faixa etária/sexo/bairro, ou seguir algoritmos mais sofisticados que consideram múltiplas causas no atestado.
 - **Heterogeneidade:** A qualidade da certificação médica pode variar enormemente entre bairros. Hospitais em áreas mais ricas podem usar menos CGs do que postos de saúde ou serviços em áreas mais pobres. A proporção de CGs que realmente eram COVID-19 também pode variar. A análise precisa ser feita bairro a bairro (ou por grupos homogêneos), considerando as características locais do sistema de saúde e da população.
- **Literatura relevante:**
 - **Naghavi, M., e outros. (2010).** *Algoritmos para melhorar a utilidade dos dados nacionais sobre causas de morte para a saúde pública.* (Discute métodos de redistribuição).
 - **Lista de causas e códigos de lixo do GBD:** O Institute for Health Metrics and Evaluation (IHME) publica listas e métodos. (www.healthdata.org)
 - **França, E., et al. (2020).** *Mudanças na qualidade das estatísticas de causas de morte no Brasil: códigos lixo entre óbitos registrados no período de 1996 a 2016.* (Contexto brasileiro).
 - **QUEM. (Várias edições).** *Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde (CID).*

Fase 3: Redistribuição de Causas Mal Definidas (CMD)

- **Objetivo intuitivo:** Similar aos CG, mas para causas ainda mais vagas, como "febre" ou "senilidade" listadas como causa básica (Capítulo XVIII da CID-10, códigos R). Tentar adivinhar quantos desses eram COVID-19.
- **Exemplo:** Na festa, algumas pessoas saíram e anotaram apenas "sintomas gerais". Se você notar que nos grupos onde mais gente anotou "sintomas gerais" também havia mais gente saindo por "intoxicação alimentar" confirmada, pode suspeitar que parte dos "sintomas gerais" era, na verdade, a mesma intoxicação.
- **Componente Teórico (Simplificado):** CMDs (geralmente códigos R00-R99 da CID-10) são ainda menos informativos que CGs. O artigo menciona o **Método de Ledermann**. A ideia central (simplificada) é usar regressão estatística:
 1. Calcular para *proporção* de óbitos por CMD em cada bairro (ou área).
 2. Calcular para *proporção* de óbitos por COVID-19 em cada bairro.
 3. Verificar se há uma correlação estatística: bairros com alta proporção de CMDs também tendem a ter alta (ou baixa) proporção de COVID-19?
 4. Se houver uma correlação positiva significativa, a inclinação da reta de regressão (o coeficiente "beta") pode ser usada para estimar qual fração dos óbitos por CMD deveria ser redistribuída para COVID-19.
- **Como o Script Faz (Simulação):**
 - Filtra óbitos por "Mal Definidas" e "COVID-19".
 - Calcula as proporções (Prop_CMD, Prop_COVID) para cada bairro/sexo (embora a regressão real não seja feita).
 - Simula um *coeficiente de redistribuição* (Coeficiente_CMD_Redistribuído).
 - Calcula a quantidade a ser redistribuída (CMD_redistribuído_para_COVID) multiplicando os CMDs observados por esse coeficiente simulado.
- **Replicando em Salvador (Heterogeneidade):**
 - **Dados Reais:** Precisa dos dados do SIM com códigos CID-10.
 - **Método Real:**

- **Número de áreas:** O método de Ledermann (regressão) precisa de um número razoável de unidades de análise (bairros) com dados estáveis. Bairros com pouquíssimos óbitos podem precisar ser agrupados com vizinhos similares socioeconomicamente. 50 bairros (como no script corrigido) começa a ser um número mais adequado para uma regressão do que apenas 5.
- **Executando a Regressão:** Usar a função `lm()` do R para ajustar o modelo (`lm(Prop_COVID ~ Prop_CMD, data = dados_por_bairro)`), possivelmente estratificando por sexo ou grupos etários. O coeficiente da `Prop_CMD` seria o "beta" estimado.
- **Heterogeneidade:** A prevalência de CMDs pode variar muito com o acesso a diagnóstico e a qualidade do preenchimento dos atestados. A relação estatística (correlação) entre CMD e COVID-19 pode ser diferente em grupos de bairros distintos. Pode ser necessário rodar regressões separadas para diferentes estratos da cidade ou incluir variáveis socioeconômicas no modelo.
- **Literatura relevante:**
 - **Ledermann, S. (1955).** *Distribuição de mortes de causa indeterminada*. (Artigo original, em francês).
 - **Nações Unidas. (1983).** *Manual X*. (Descreve o método).
 - **Mello Jorge, MHP, e outros. (2008).** *A mortalidade de idosos no Brasil: a questão das causas mal definidas*. (Contexto brasileiro sobre CMDs).

Fase 4: Cálculo Final dos Óbitos Esperados

- **Objetivo intuitivo:** Juntar todas as peças: os óbitos originalmente registrados como COVID (corrigidos pela estimativa de 'perdidos') + os óbitos reclassificados dos Códigos Garbage + os óbitos reclassificados das Causas Mal Definidas.
- **Exemplo:** O número final de convidados da festa é: (os que você contou x fator de correção para os que você perdeu de vista) + (os que anotaram "mal-estar" e você reclassificou) + (os que anotaram "sintomas gerais" e você reclassificou).

- **Componente Teórico (Simplificado):** É a aplicação direta da fórmula principal do estudo:

$$\text{Óbitos Esperados COVID} = (f * \text{Óbitos Observados COVID}) + \text{Óbitos Redistribuídos de CG} + \text{Óbitos Redistribuídos de CMD}$$

Onde f é o fator de sub-registro, Óbitos Observados COVID é o número inicial nos registros, e os outros dois termos são as quantidades calculadas nas fases 2 e 3.
- **Como o Script Faz:**
 - `cervojunção_à_esquerda` para juntar os resultados das fases anteriores (o fator (f) ou `Obitos_COVID_Corrigidos_Subregistro,CG_redistribuído_para_COVID,CMD_redistribuído_para_COVID`) para cada bairro/sexo.
 - Soma esses componentes para obter Mortes esperadas por COVID.
 - Calcula a diferença e o aumento percentual em relação ao número original observado.
- **Replicando em Salvador (Heterogeneidade):**
 - Este passo é a aplicação final da fórmula, usando os valores *reaisdef*, `CG_j`, e `eCMD_j` que você calculou/estimou *especificamente para cada bairro* (ou grupo de bairros) nas fases anteriores.
 - O resultado será um número estimado de óbitos por COVID-19 para cada bairro, refletindo as correções aplicadas com base nas suas características e na qualidade dos dados locais. A soma desses resultados por bairro dará o total estimado para a cidade, que provavelmente será significativamente maior que o total originalmente registrado.

Considerações Finais para Aplicação Real:

1. **Especialização:** A aplicação rigorosa desses métodos (BGC, Ledermann, algoritmos de CG) exige conhecimento em demografia e/ou epidemiologia estatística. Colaborar com especialistas é altamente recomendável.
2. **Qualidade dos Dados:** A qualidade dos dados de população (IBGE) e mortalidade (SIM) é crucial. Verifique a consistência, cobertura e possíveis vieses nesses dados antes de aplicar os métodos.
3. **Escolha dos Métodos:** Existem variações e alternativas para cada passo (diferentes métodos para sub-registro, diferentes algoritmos para CG). A

escolha deve ser justificada e baseada na disponibilidade de dados e no contexto local.

4. **Validação e Sensibilidade:** É importante realizar análises de sensibilidade (o que acontece se você mudar um pouco as premissas, como a lista de CGs ou o fator f ?) e, se possível, validar os resultados com outras fontes de informação (ex: dados de excesso de mortalidade geral, dados de hospitalização por SRAG).
5. **Interpretação:** Os resultados são *estimativas* com incertezas. Comunique-os com clareza, explicando as fontes de incerteza e as limitações da metodologia.

Seu esforço em entender e aplicar essa metodologia é louvável. Começar com o script de simulação e entender cada passo conceitualmente é o caminho certo antes de mergulhar nos dados reais e na complexidade dos métodos demográficos/estatísticos. Boa sorte com sua análise para Salvador!