1 - Introdução.

Prof. Vinícius D. Mayrink
EST171 - Estatística Espacial

Sala: 4073 Email: vdm@est.ufmg.br

1^o semestre de 2024

Introdução: Modelos e dados espaciais.

Dados espaciais e espaço-temporais são bastante comuns.

Aparecem com frequência na TV e Internet (exemplo: previsão do tempo). Podem ser encontrados também nos aparelhos GPS, aplicativos de celular e em simples mapas reproduzidos em papel.

Construir um mapa que seja útil para expressar uma informação, e que não distorça os dados, pode não ser uma tarefa simples.

Além de criar um mapa, a Estatística Espacial se preocupa com questões que não podem ser diretamente respondidas olhando para esse mapa.

Estas questões referem-se a processos que supomos estar gerando os dados observados.

A Inferência Estatística para esses processos é muitas vezes desafiadora, mas ela é necessária para tirar conclusões sobre questões que nos interessam.

Introdução: Modelos e dados espaciais.

Pesquisadores de diversas áreas (climatologia, ecologia, saúde, mercado imobiliário e outras) deparam-se constantemente com a tarefa de analisar dados que:

- são altamente multivariados
  (com muitas covariáveis e variáveis respostas);
- apresentam referência geográfica (representação via mapas);
- são correlacionados no tempo
  (longitudinalmente ou seguindo outra estrutura temporal).

Isso motiva a construção de modelos hierárquicos para lidar com essa estrutura espacial ou espaço-temporal.

Exemplos:

- Investigação epidemiológica: estudar as taxas de câncer de pulmão e mama por município e ano em um certo estado. Considere as variáveis "fumante", "mamografia" e outras informações importantes sobre o estágio da doença.

- Investigação epidemiológica: avaliar o padrão espacial da incidência de uma doença para determinar se existe um cluster de incidências. Além disso, podemos querer saber se há fatores relacionados ao cluster como idade, nível de pobreza e fontes poluidoras.

- Investigação meteorológica: avaliar dados de temperatura e precipitação medidos, por hora ou por dia, em uma rede de estações de monitoramento que podem estar localizadas em diferentes elevações do terreno ou com alguma tendência de localização.

Podemos estar interessados em simplesmente visualizar os dados coletados mas, em geral, também desejamos realizar procedimentos de inferência tais como:

- modelagem de tendências;
- estimação da estrutura de correlação dos parâmetros do modelo;
- teste de hipóteses ou comparação de modelos competidores;
- previsão de observações em tempos ou locais não observados.

Um livro importante e considerado um marco em estatística espacial é Cressie (1993). Entretanto, esta referência é antiga e não aborda desenvolvimentos recentes da área, principalmente aqueles ligados a aspectos computacionais que permitem trabalhar com modelos hierárquicos espaciais sofisticados.

Foco principal: modelagem, computação e análise de dados.

No foco computacional, iremos utilizar ferramentas disponíveis no *software R*.

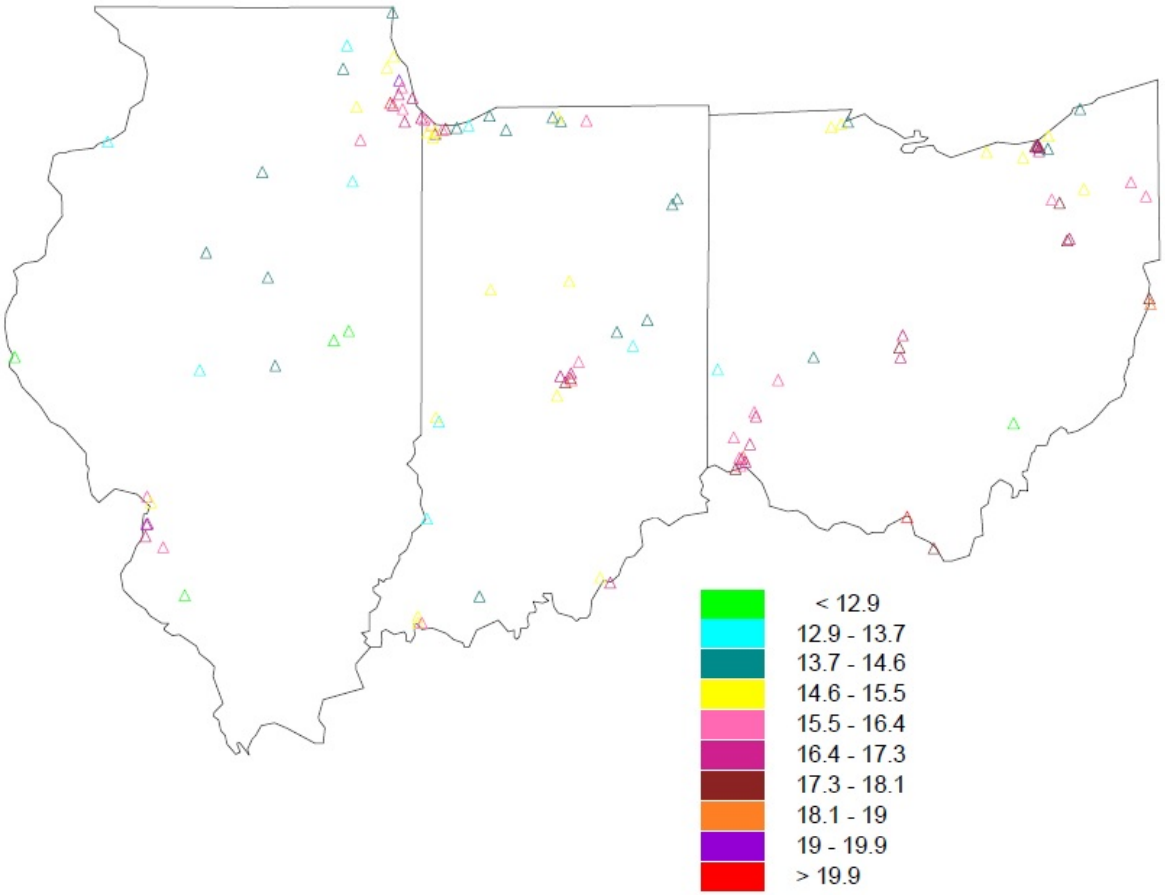
Classificamos os dados espaciais em três tipos. Considere $Y(\mathbf{s})$ um vetor aleatório relacionado ao local $\mathbf{s} \in \mathbb{R}^r$.

Dados com referência pontual (point-referenced data, geostatistical): $\mathbf{s}$ varia continuamente em $D \subset \mathbb{R}^r$ fixo e contendo um retângulo $r-$dimensional de volume positivo.

Dados de área (areal data): $\mathbf{s}$ varia continuamente em $D \subset \mathbb{R}^r$ fixo (com formato regular ou irregular) e particionado em um número finito de unidades de área com fronteiras bem definidas.

Dados pontuais com padrão (point pattern data): $\mathbf{s}$ varia continuamente em $D \subset \mathbb{R}^r$ aleatório. O conjunto de índices em D identifica os locais dos eventos aleatórios que formam o padrão espacial de pontos. Poderíamos ter: $Y(\mathbf{s}) = 1$ $\forall \mathbf{s} \in D$ (indicando a ocorrência do evento) e $Y(\mathbf{s}) = 0$ caso contrário. Podemos também adicionar informação de covariáveis produzindo um processo pontual com padrão marcado.

Exemplo - Dados com referência pontual (geoestatístico): 114 locais de monitoramento da poluição do ar em Illinois, Indiana e Ohio. A variável de interesse é a média anual (ano 2001) do nível PM2.5 = *Particle Matter less than 2.5 microns* medida (em *ppb*) em cada local.



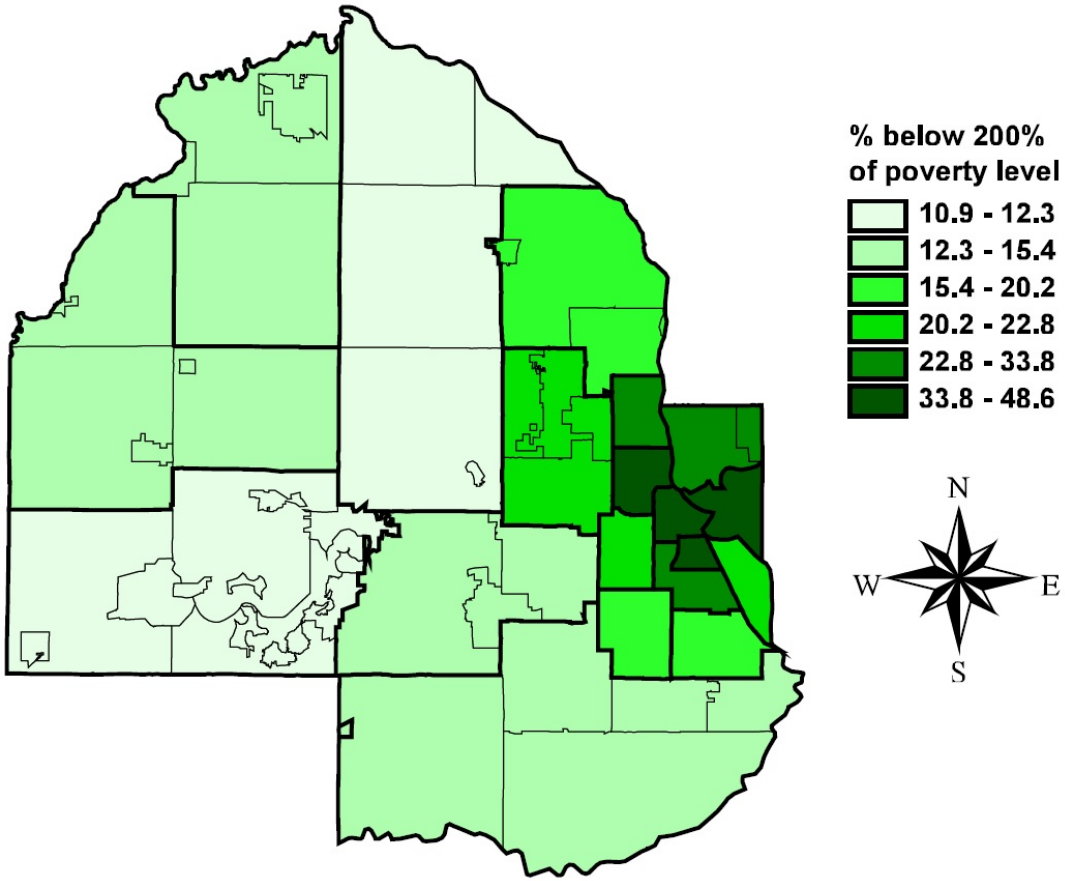
Exemplo - Dados com referência pontual (geoestatístico): Monitoramento da poluição do ar.

Podemos estar interessados em um modelo para a distribuição geográfica destes níveis de PM2.5 que avalie a correlação espacial.

Possíveis covariáveis que poderíamos usar:
"industrialização local", "intensidade de tráfego", etc.

Métodos tradicionais para análise deste tipo de dado serão apresentados mais adiante.

Exemplo - Dados de área (lattice): Porcentagem da população amostrada com renda familiar abaixo de 200% do limite federal (EUA) de pobreza. As unidades de área são as unidades regionais de amostragem definidas no condado de Hennepin/Minnesota.



Exemplo - Dados de área (lattice):

- A figura anterior é um exemplo de um *choropleth map*. Ele usa tonalidades de cores (ou de uma cor, ex. escala de cinza) para classificar valores em algumas classes; ideia do histograma.
- A partir do *choropleth map* sabemos quais regiões são adjacentes (fazem fronteira) umas com as outras.
- Os locais $\mathbf{s} \in D$ são, neste caso, as regiões (ou blocos) observadas no mapa, as quais denotamos por B_i com $i = 1, \ldots, n$ (distinguindo de $\mathbf{s_i}$).
- Pode ser útil pensar nos centroides de cada região formando os vértices de uma lattice irregular cujos pontos são conectados se e somente se as regiões correspondentes são "vizinhas" no mapa.

Exemplo - Dados de área (lattice):

A nomenclatura "lattice" pode ser enganosa visto que faz referência mais direta a observações organizadas em um "grid" do tipo tabuleiro de xadrez. Realmente existem dados organizados desta maneira (ex.: amostras de campos de agricultura, pixels em uma imagem); neste caso temos a estrutura de uma *lattice regular*.

Na prática a maioria dos dados de área são observações sobre uma *lattice irregular* como em uma coleção de municípios ou outra configuração de fronteiras.

Exemplo - Dados pontuais e de área (misaligned data): Alguns dados espaciais apresentam observações pontuais e de área, requerendo uma análise simultânea. A figura a seguir, mostra as fronteiras de "zip code" (CEP) na região metropolitana de Atlanta/GA e o nível máximo (*ppm*) de ozônio em um período de 8 horas medido em 10 estações de monitoramento em 15/Julho/1995.

As observações foram coletadas em estações fixas com coordenadas espaciais (ex.: latitude e longitude) conhecidas.

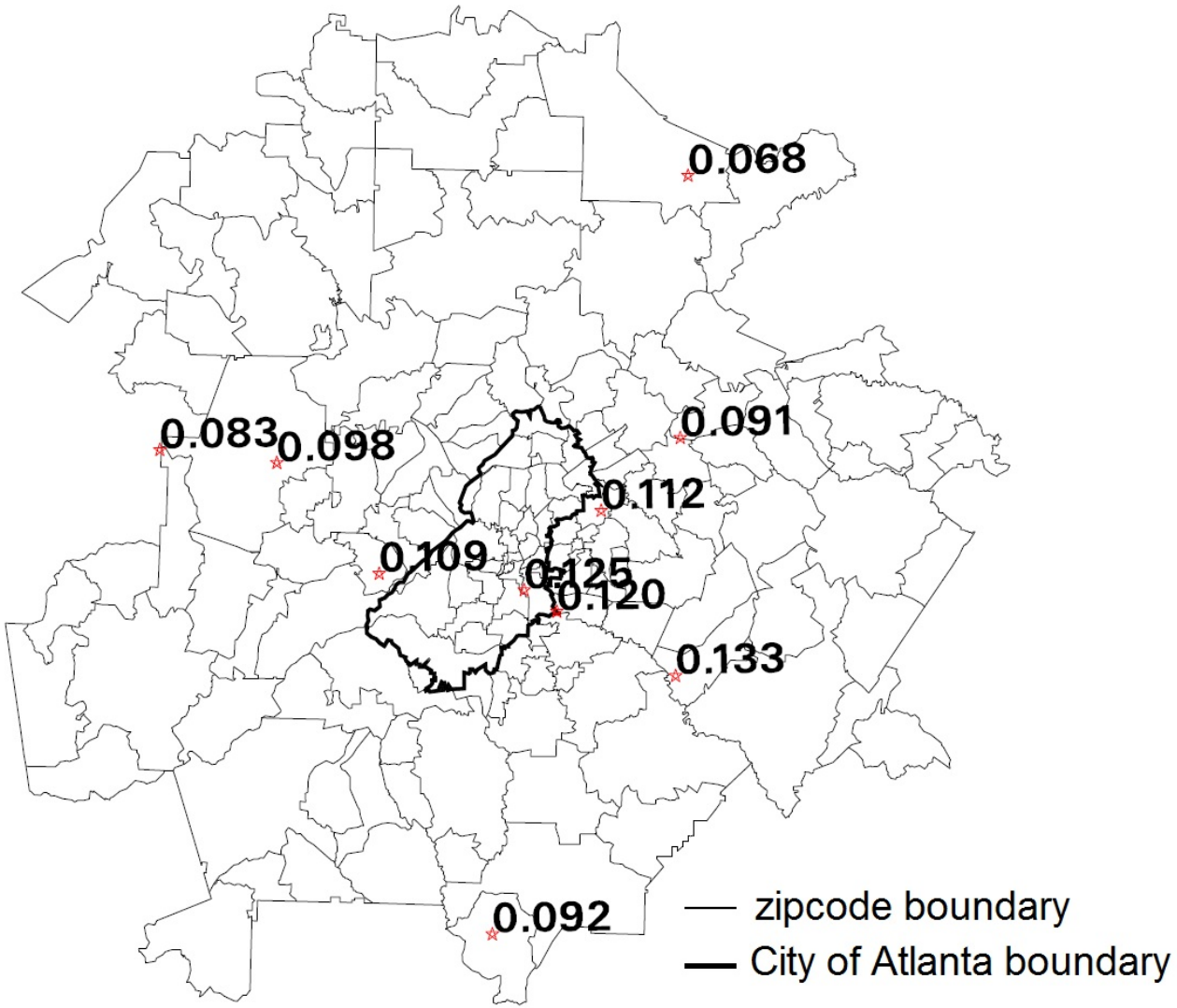
Assumimos que $Y(\mathbf{s_i})$ é aleatório, mas $\mathbf{s_i}$ é fixo para $i = 1, \ldots, 10$.

Um segundo elemento nestes dados é o número de crianças na área de um *zip code* que reportaram, no Pronto Atendimento local, sintomas de Asma.

Não temos as coordenadas do local exato onde vive a criança (confidencialidade).

Questão: Podemos estabelecer uma conexão entre altos níveis de ozônio e o maior número de visitas de crianças com sintomas de Asma? Uma investigação estatística formal requer um *realignment* preliminar destes dados.

Exemplo - Dados pontuais e de área (misaligned data):



Exemplos - Dados de um processo pontual espacial: residências de pessoas sofrendo de uma particular doença, locais de certas espécies de árvore em uma floresta, locais de um município onde novas residências são construídas.

A resposta Y é frequentemente fixa (ocorrência ou não do evento de interesse). Os locais $\mathbf{s_i}$ são tratados como aleatórios.

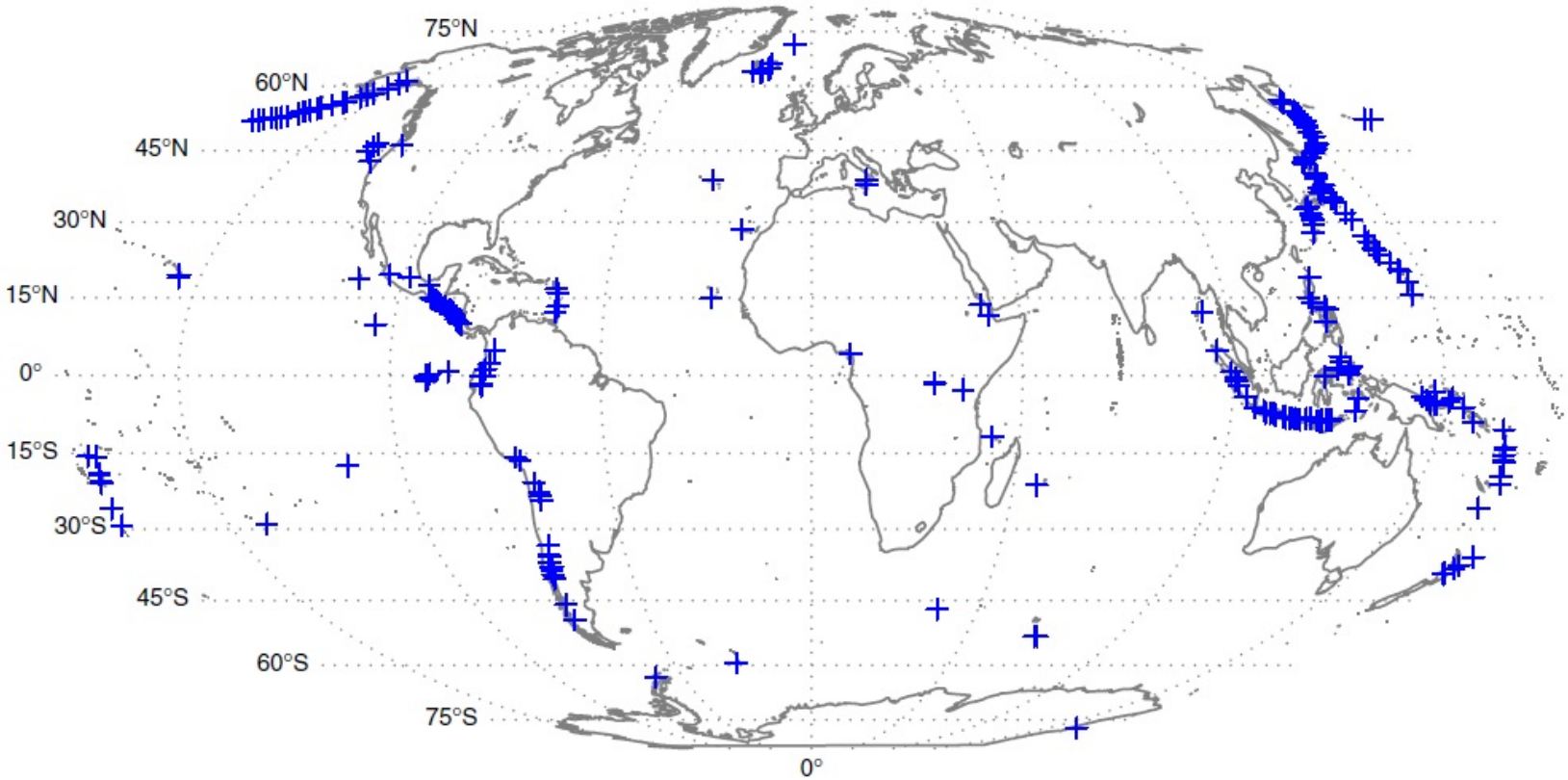
Estudos de conglomerados (*clustering*) de eventos, visando determinar:

- se os pontos tendem a estar espacialmente próximos uns dos outros;
- se os pontos se encontram independentemente e homogeneamente posicionados no espaço em um processo aleatório sem conglomerados.

Em contraste com dados de área, aqui (e também para dados com referência pontual) os locais precisos devem ser conhecidos. Estes dados, em geral, são protegidos (*encrypted*) para preservar a privacidade dos indivíduos amostrados.

Fundamentos de cartografia.

Dados espaciais apresentam uma referência espacial: valores de coordenadas e um sistema de referência para estas coordenadas. Exemplo: considere a localização de vulcões na Terra.



Poderíamos listar as coordenadas de todos os vulcões conhecidos através de pares de valores (longitude, latitude) medindo o ângulo em relação à longitude zero (Meridiano de Greenwich) e à latitude zero (Linha do Equador).

Este sistema de referência é conhecido como Sistema Geodésico Mundial (*World Geodetic System - WGS84*). Ele é utilizado em aparelhos de GPS.

A Terra é redonda!
Então (longitude, latitude) $\neq (x,y) =$ coordenadas em um plano.

A projeção de um mapa é uma representação sistemática de toda ou parte da superfície terrestre em um plano.

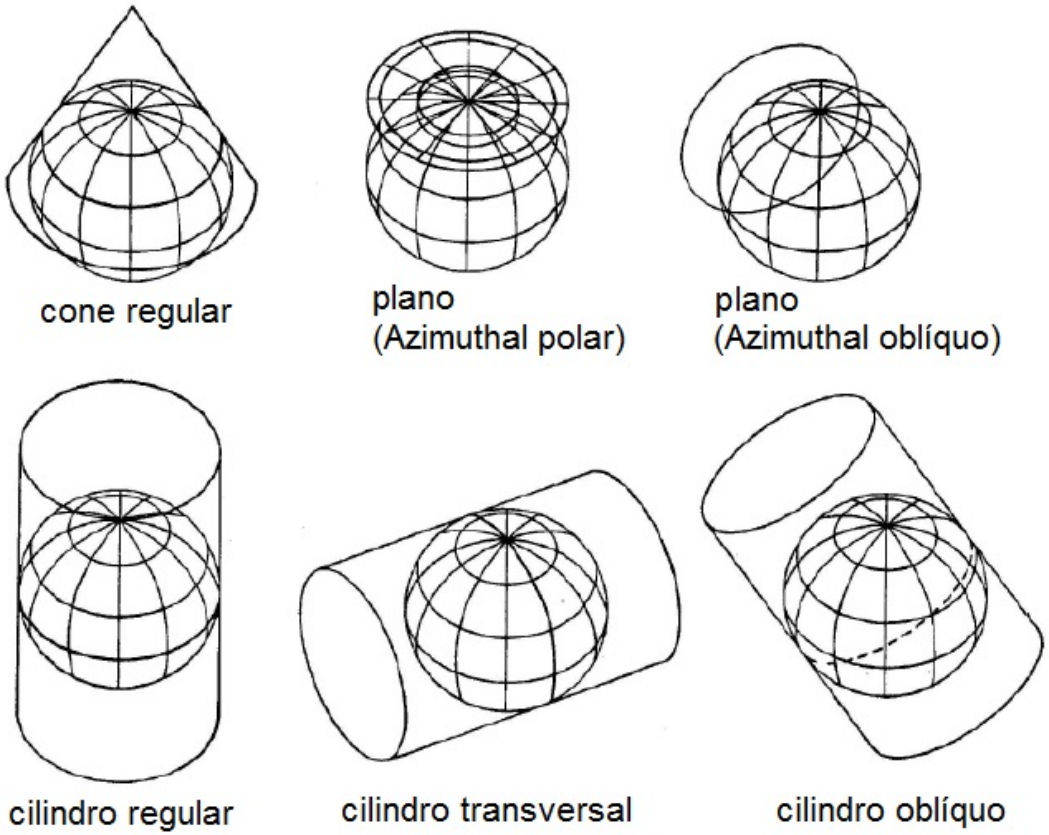
Fato conhecido em topologia: Uma esfera não pode ser transformada em um plano sem que ocorram distorções.

Um cartógrafo precisa escolher a(s) característica(s) que deseja mostrar com maior precisão no mapa (não existe a "melhor" projeção de um mapa).

Estratégia: usar uma superfície intermediária para fazer a transição da esfera para o plano; a esfera é primeiramente projetada em uma superfície intermediária.

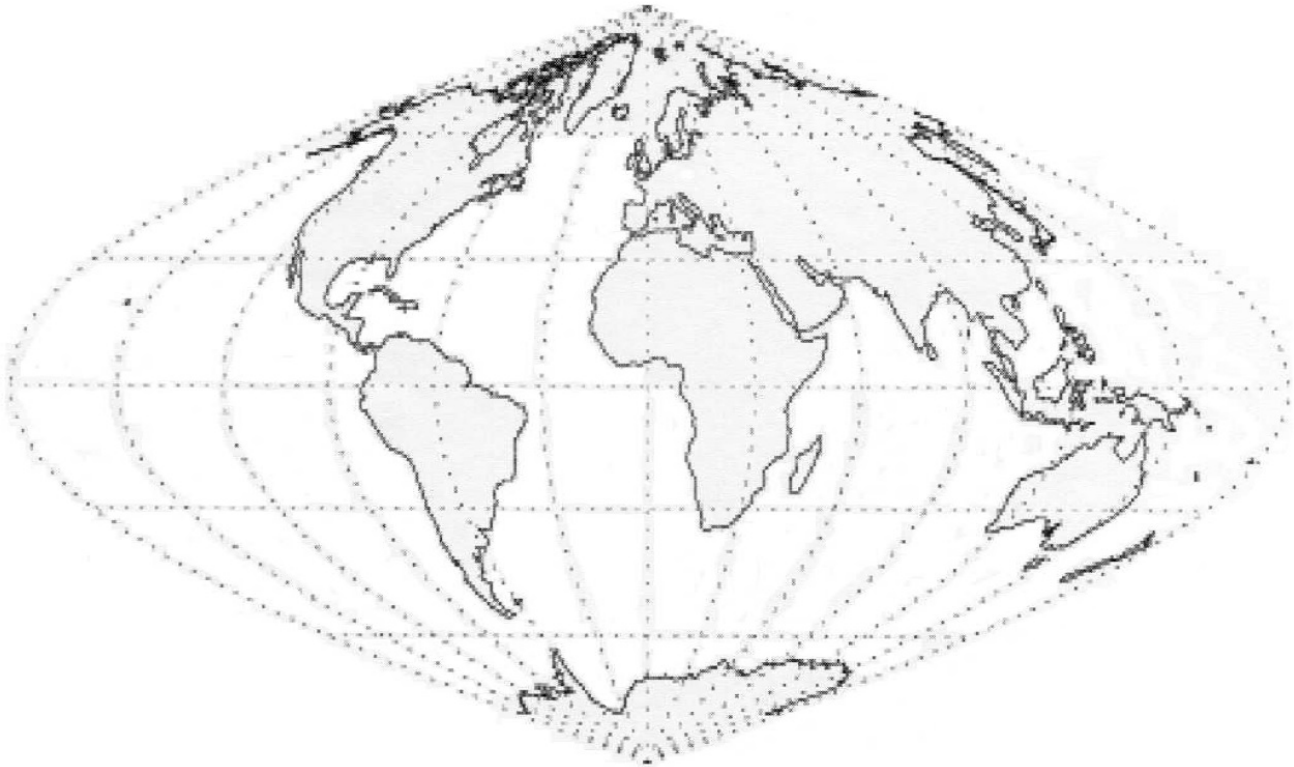
Superfícies mais usadas para transição: o cilindro, o cone e o próprio plano.
Diferentes orientações destas superfícies levam a diferentes classes de projeções.

Nenhuma projeção é capaz de preservar as distâncias originais.
É possível preservar as áreas originais.



Notação: $(\lambda,\theta) =$ (longitude, latitude).
Ideia básica para determinar equações de projeção de mapas:

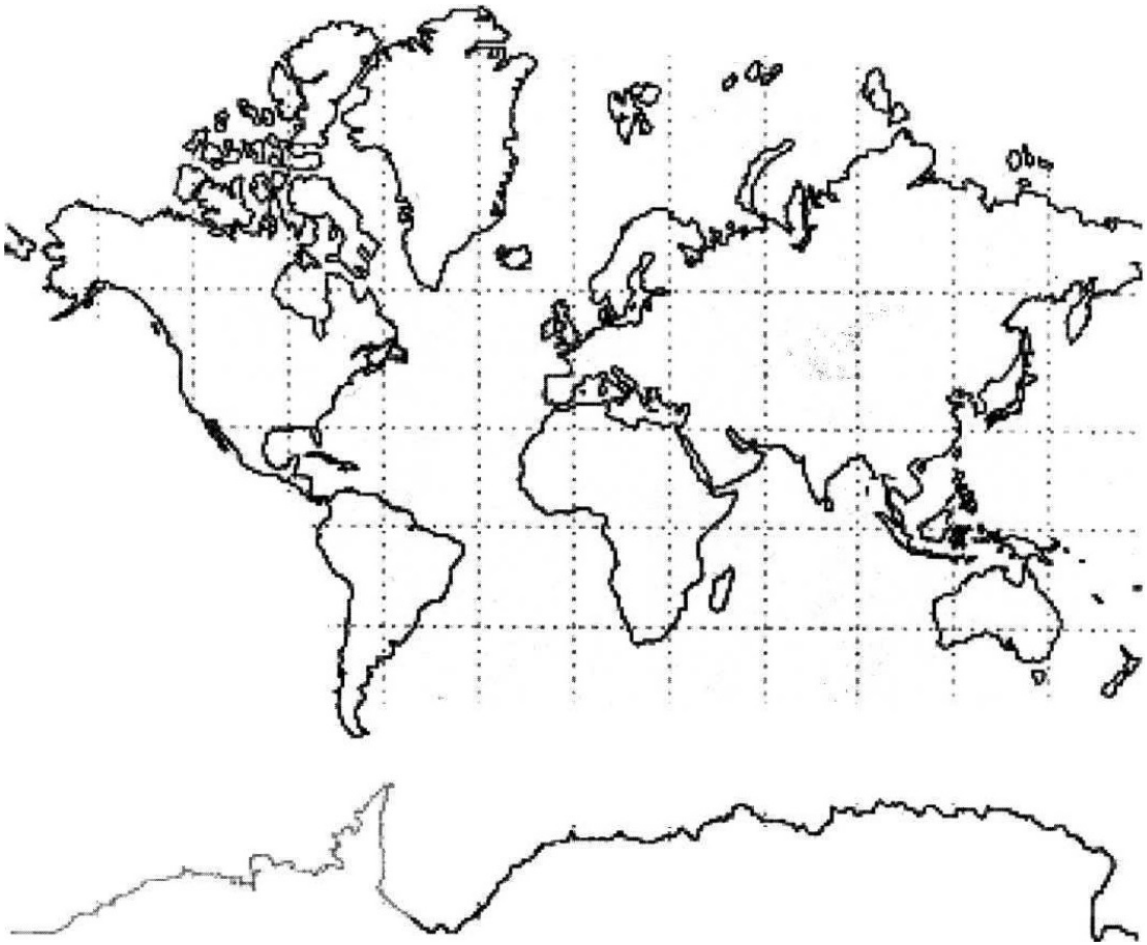
Partindo de uma esfera com sistema (λ,θ), construa um novo sistema de coordenadas (x,y) tal que: $\quad x = f(\lambda,\theta) \quad$ e $\quad y = g(\lambda,\theta)$.
A escolha de f e g baseia-se em propriedades desejadas para o mapa.



Projeção sinusoidal (preserva a área das regiões).

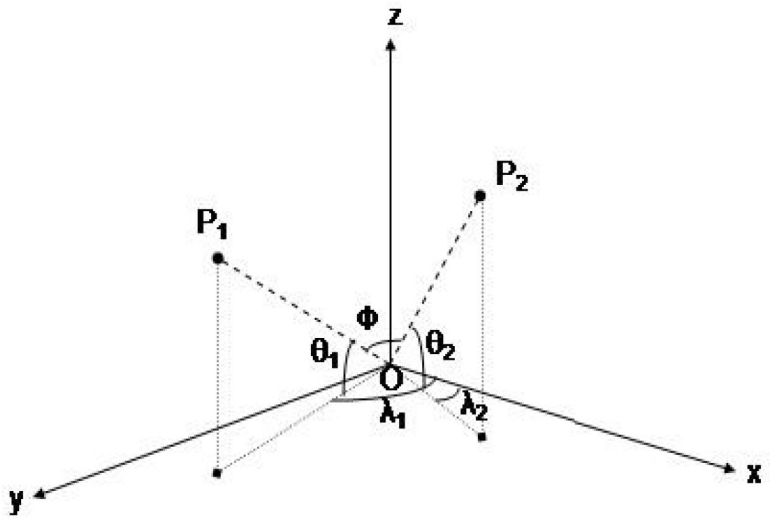
Apresenta linhas retas igualmente espaçadas representando os paralelos e uma reta vertical para o meridiano central.

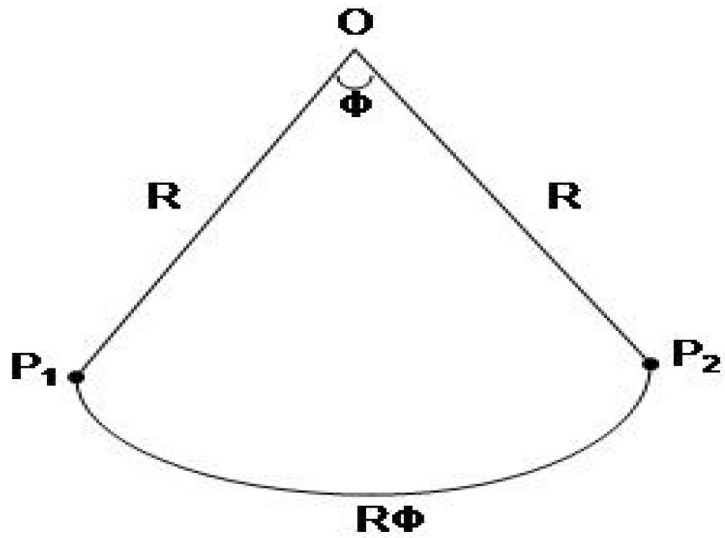
Neste caso, temos $f(\lambda, \theta) = R\lambda\cos(\theta)$ e $g(\lambda, \theta) = R\theta$, sendo R o raio da Terra.



Projeção de Mercator (não preserva a área das regiões).
Mercator distorce bastante as regiões próximas aos polos.
Os meridianos são todos representados por retas paralelas.
Neste caso, temos: $f(\lambda, \theta) = R\,\lambda$ e $g(\lambda, \theta) = R\,\ln[\tan(\pi/4 + \theta/2)]$.

Fundamentos de cartografia: cálculo da distância geodésica.





Considere dois pontos na superfície da Terra: $P_1 = (\lambda_1, \theta_1)$ e $P_2 = (\lambda_2, \theta_2)$ sendo λ_i = longitude do ponto i e θ_i = latitude do ponto i.

Distância geodésica: $D = R\phi$ é comprimento do arco entre P_1 e P_2;
R = raio da Terra e ϕ = ângulo (em radianos; ex.: 0, $\pi/2$, π, $3\pi/2$, 2π) entre as retas ligando P_1 e P_2 ao centro $\mathbf{O} = (0,0)$.

Relações entre as coordenadas (x,y,z) e (λ,θ) via trigonometria:

$$x = R\ \cos(\theta)\ \cos(\lambda) \qquad y = R\ \cos(\theta)\,\text{sen}(\lambda) \qquad z = R\,\text{sen}(\theta)$$

Sejam $\mathbf{u_1} = (x_1, y_1, z_1)'$ e $\mathbf{u_2} = (x_2, y_2, z_2)'$.

O ângulo entre os vetores não nulos $\mathbf{u_1}$ e $\mathbf{u_2}$ é igual a $\arccos\left\{\frac{\langle \mathbf{u_1},\mathbf{u_2}\rangle}{\|\mathbf{u_1}\|\|\mathbf{u_2}\|}\right\}$,
logo $\cos(\phi) = \frac{\langle \mathbf{u_1},\mathbf{u_2}\rangle}{\|\mathbf{u_1}\|\|\mathbf{u_2}\|}$, sendo $\langle \mathbf{u_1}, \mathbf{u_2}\rangle = \mathbf{u_1}'\mathbf{u_2}$ e $\|\mathbf{u_1}\| = \sqrt{x_1^2 + y_1^2 + z_1^2}$.

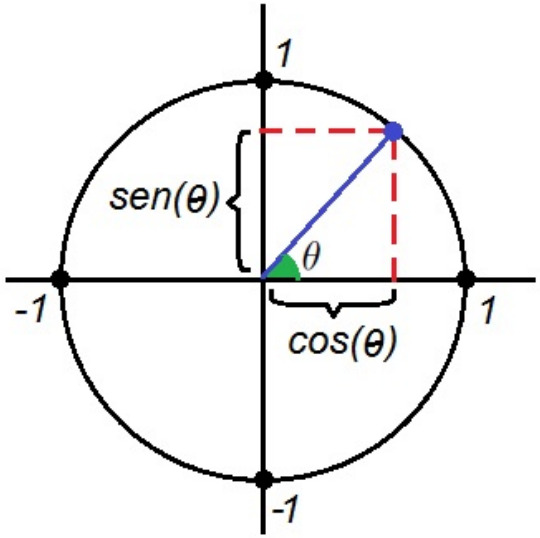
Calculamos:

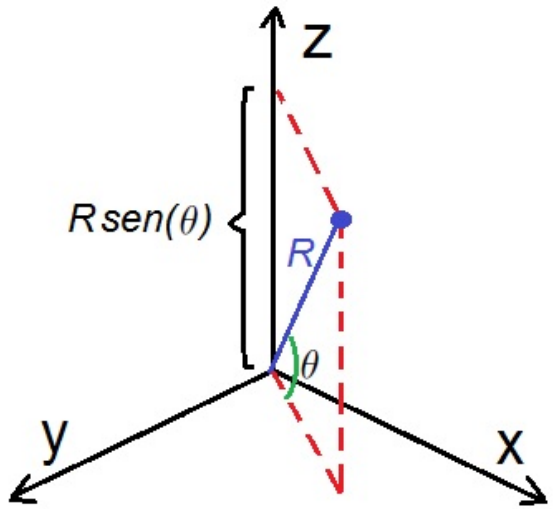
$$\begin{aligned}
\langle \mathbf{u_1}, \mathbf{u_2}\rangle &= R^2\ \cos(\theta_1)\cos(\lambda_1)\cos(\theta_2)\cos(\lambda_2) + \\
&\quad + R^2\ \cos(\theta_1)\text{sen}(\lambda_1)\cos(\theta_2)\text{sen}(\lambda_2) + R^2\,\text{sen}(\theta_1)\text{sen}(\theta_2) \\
&= R^2\left[\cos(\theta_1)\cos(\theta_2)\cos(\lambda_1 - \lambda_2) + \text{sen}(\theta_1)\text{sen}(\theta_2)\right]
\end{aligned}$$

visto que $\|\mathbf{u_1}\| = \|\mathbf{u_2}\| = R$, temos

$$D = R\phi = R\ \arccos\left[\cos(\theta_1)\cos(\theta_2)\cos(\lambda_1 - \lambda_2) + \text{sen}(\theta_1)\text{sen}(\theta_2)\right].$$

Mais detalhes sobre os resultados acima a seguir...





$$
\begin{aligned}
R &= \sqrt{x^2 + y^2 + z^2} \\
R^2 &= x^2 + y^2 + z^2 \\
x^2 + y^2 &= R^2 - z^2 \\
x^2 + y^2 &= R^2 - R^2 sen^2(\theta) \\
x^2 + y^2 &= R^2[1 - sen^2(\theta)] \\
x^2 + y^2 &= R^2\, cos^2(\theta)
\end{aligned}
$$

$$
\begin{aligned}
& R^2\, cos^2(\theta) = \\
&= R^2\, cos^2(\theta)[cos^2(\lambda) + sen^2(\lambda)] \\
&= R^2\, cos^2(\theta)cos^2(\lambda) + R^2\, cos^2(\theta)sen^2(\lambda) \\
&= [R\, cos(\theta)cos(\lambda)]^2 + [R\, cos(\theta)sen(\lambda)]^2
\end{aligned}
$$

Conclusão:

$x = R\, cos(\theta)cos(\lambda)$,

$y = R\, cos(\theta)sen(\lambda)$.

$$
\begin{aligned}
\langle \mathbf{u_1}, \mathbf{u_2} \rangle &= \mathbf{u_1}'\mathbf{u_2} = x_1x_2 + y_1y_2 + z_1z_2 \\
&= R\cos(\theta_1)\cos(\lambda_1)\, R\cos(\theta_2)\cos(\lambda_2) + R\cos(\theta_1)\mathrm{sen}(\lambda_1)\, R\cos(\theta_2)\mathrm{sen}(\lambda_2) + \\
&\quad + R\mathrm{sen}(\theta_1)\, R\mathrm{sen}(\theta_2) \\
&= R^2[\cos(\theta_1)\cos(\lambda_1)\cos(\theta_2)\cos(\lambda_2) + \cos(\theta_1)\mathrm{sen}(\lambda_1)\cos(\theta_2)\mathrm{sen}(\lambda_2) + \\
&\quad + \mathrm{sen}(\theta_1)\,\mathrm{sen}(\theta_2)] \\
&= R^2\{\cos(\theta_1)\cos(\theta_2)[\cos(\lambda_1)\cos(\lambda_2) + \mathrm{sen}(\lambda_1)\mathrm{sen}(\lambda_2)] + \mathrm{sen}(\theta_1)\,\mathrm{sen}(\theta_2)\} \\
&= R^2\{\cos(\theta_1)\cos(\theta_2)[\cos(\lambda_1 - \lambda_2)] + \mathrm{sen}(\theta_1)\,\mathrm{sen}(\theta_2)\}
\end{aligned}
$$

Revisando duas definições a serem usadas com frequência.
Sejam os vetores $\mathbf{u} = (u_1, \ldots, u_d)'$ e $\mathbf{v} = (v_1, \ldots, v_d)' \in \mathbb{R}^d$.

Distância Euclidiana entre $\mathbf{u}$ e $\mathbf{v}$: comprimento do segmento conectando os pontos $\mathbf{u}$ e $\mathbf{v}$.

$$\|\mathbf{u} - \mathbf{v}\| = \sqrt{(u_1 - v_1)^2 + \ldots + (u_d - v_d)^2}.$$

Norma Euclidiana de $\mathbf{u}$: medida do comprimento do vetor u = distância de $\mathbf{u}$ até a origem $\mathbf{0}$.

$$\|\mathbf{u}\| = \sqrt{u_1^2 + \ldots + u_d^2}.$$