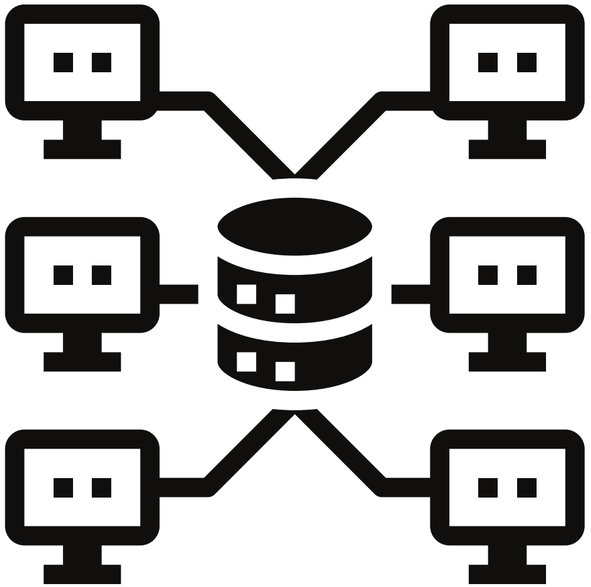
Data Science e Statistical Computing

Bases de Dados



Prof. Tiago Marum
FIAP e THM Estatística
@thmestatistica

ESTUDO DE CASO - VENDAS DE UM E-COMMERCE

ANÁLISE DE DADOS DE VENDAS DE UM E-COMMERCE

Uma visão detalhada sobre os dados de pedidos, produtos e entregas

- O dataset atual registra informações detalhadas sobre **pedidos de um e-commerce**, incluindo valores, categorias de produtos, status de entrega e dados logísticos.
- Ele permite analisar o **desempenho das vendas**, identificar padrões de compra e avaliar a eficiência na distribuição.
- Com esses dados, é possível **detectar gargalos logísticos**, entender fatores que levam a **cancelamentos** e **otimizar estratégias de vendas e frete**, tornando-se uma ferramenta essencial para melhorar a gestão no comércio eletrônico.

E-Commerce Sales Dataset

Analyzing and Maximizing Online Business Performance

k kaggle.com

ANÁLISE DE DADOS DE VENDAS DE UM E-COMMERCE

Uma visão detalhada sobre os dados de pedidos, produtos e entregas

- **Order ID**: Identificador único para cada pedido realizado.
- **Date**: Data em que o pedido foi efetuado.
- **Status**: Situação atual do pedido (por exemplo, "Enviado", "Cancelado").
- **Fulfilment**: Responsável pelo processamento e envio do pedido (Amazon ou vendedor terceirizado).
- **Sales Channel**: Plataforma pela qual o pedido foi realizado (por exemplo, "Amazon.in").
- **Ship Service Level**: Nível de serviço de envio escolhido pelo cliente (por exemplo, "Padrão", "Expresso").
- **Style**: Descrição ou nome do modelo do produto.
- **SKU**: Unidade de Manutenção de Estoque; código único usado para identificar produtos no inventário.
- **Category**: Categoria ou segmento ao qual o produto pertence (por exemplo, "Eletrônicos", "Roupas").
- **Currency**: Moeda utilizada na transação (por exemplo, "INR" para Rúpias Indianas).
- **Amount**: Valor total do pedido na moeda especificada.
- **Ship City**: Cidade de destino para onde o pedido foi enviado.
- **Ship State**: Estado ou província de destino do pedido.
- **Ship Postal Code**: Código postal do endereço de entrega.
- **Ship Country**: País de destino do pedido.
- **Promotion IDs**: Identificadores de promoções ou descontos aplicados ao pedido.
- **B2B**: Indica se o pedido foi uma transação business-to-business (Verdadeiro ou Falso).
- **Fulfilled By**: Entidade responsável pelo cumprimento e envio do pedido (por exemplo, "Amazon", "Easy Ship").
- **Unnamed: 22**: Coluna sem nome que pode conter dados irrelevantes ou estar vazia.

PERGUNTAS/POSSIBILIDADES DE ANÁLISES ESTATÍSTICAS

Desempenho de Vendas

- Quais são os produtos mais vendidos e quais geram mais receita?
  - Identificar os produtos com maior volume de vendas e maior faturamento total
- Como estão distribuídos os valores dos pedidos?
  - O que é mais “comum”, “raro”, “baixo”, “alto”...
- Qual a porcentagem de status dos pedidos?
  - Entender a porcentagem de entregues, cancelamentos e em processamento

Logísticas e Geográficas

- Quais estados recebem mais pedidos?
- Qual o tempo médio entre a compra e a entrega, considerando diferentes tipos de envio?
- Há padrões geográficos nos cancelamentos dos pedidos?

Análise de Clientes e Estratégias de Vendas

- Os pedidos B2B (business-to-business) possuem valores médios maiores do que os pedidos normais?
  - Comparar o ticket médio de pedidos B2B vs. consumidores finais
- Promoções realmente aumentam as vendas?
- Os pedidos enviados pela própria Amazon têm uma taxa de entrega mais eficiente do que os enviados por terceiros?

Padrões e Tendências de Mercado

- Existe uma sazonalidade nas vendas?
  - Verificar se há padrões de aumento e queda nos pedidos ao longo dos meses ou dias da semana.

DESASTRES AMBIENTAIS DO ROMPIMENTO DAS BARRAGENS EM MARIANA E BRUMADINHO

Igam
Instituto Mineiro de Gestão das Águas

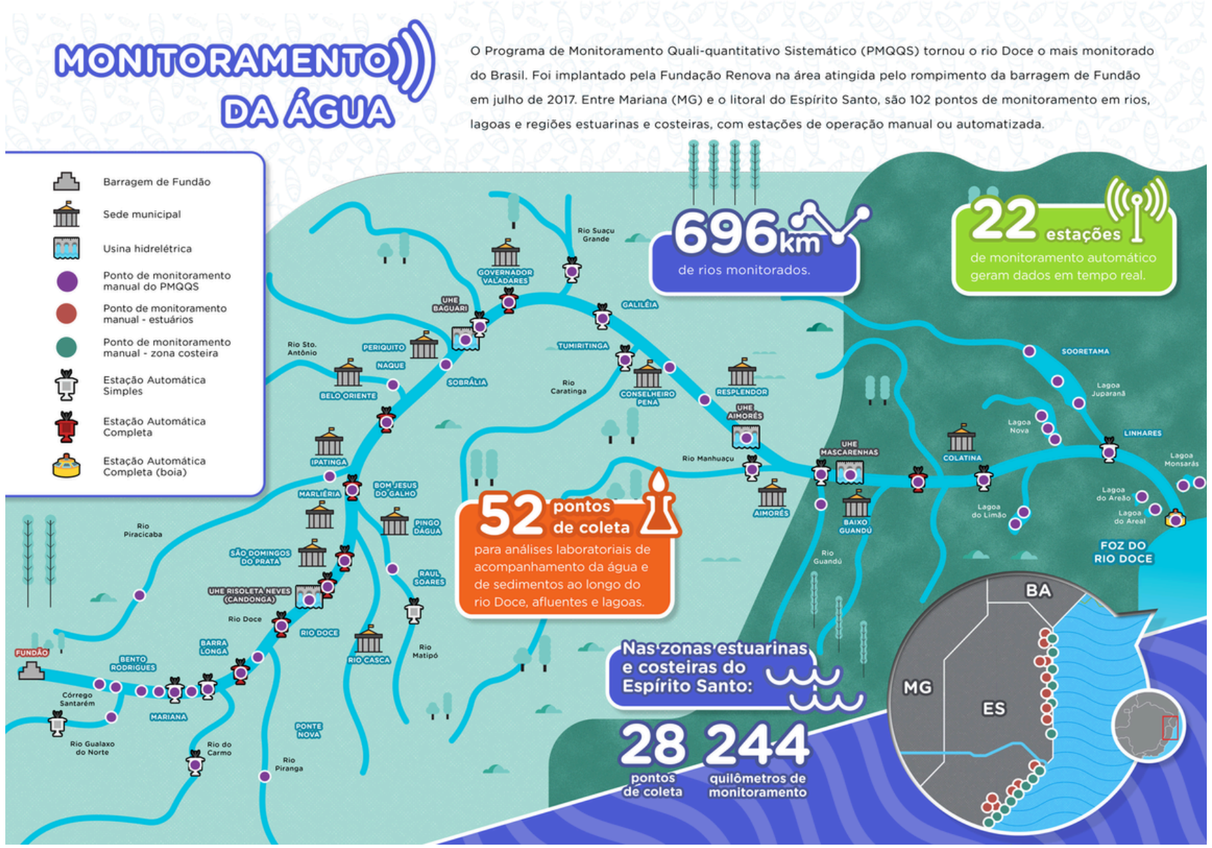
Acesse também SEMAD feam IEF

Início > Acessar Dados e Boletins de Qualidade das Águ...

Acessar Dados e Boletins de Qualidade das Águas em Minas Gerais

- Dados de Qualidade das Águas – série histórica de parâmetros de qualidade das águas e do Índice de Qualidade das Águas (IQA) de 1997 até 2020(NOVO!).
- Mapas de Qualidade das Águas Anuais – ano base 2019
- Boletins de Qualidade das Águas por UPGRH – ano base 2019

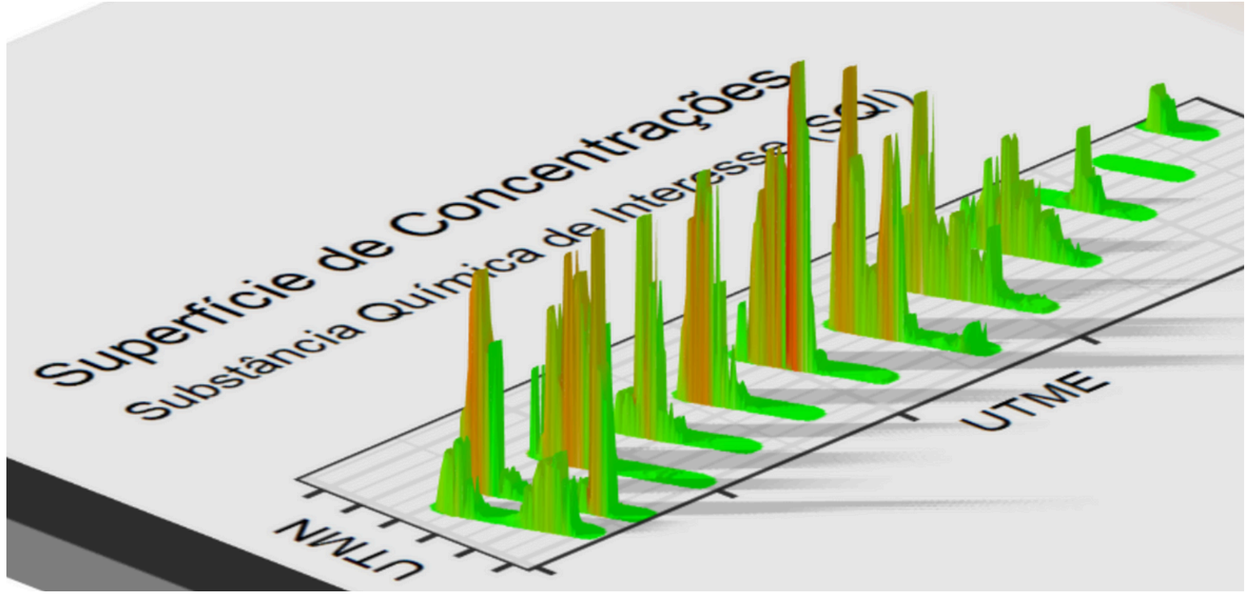
TITLE	LAST MODIFIED
BOLETINS CH 2021	9/12/22
Dados Parâmetros 2020	3/4/21



ESTUDOS DE RISCO À SAÚDE HUMANA

O estudo de risco à saúde humana avalia a probabilidade de efeitos adversos decorrentes da exposição a agentes químicos ou físicos em ambientes específicos.

$$R = \frac{C \cdot IR \cdot EF \cdot ED}{BW \cdot AT}$$



C = Concentração do contaminante

IR = Taxa de Ingestão/Inalação

EF = Frequência de Exposição

ED = Duração da Exposição

BW = Peso corporal médio

AT = Tempo de averiguação

Só a questão de "como definir o C" já é central

Outros parâmetros dependendo de coletas em entrevistas ou valores estimados por um especialista de risco e da área de estudo

ESTUDOS DE RISCO À SAÚDE HUMANA

$$R = \frac{C \cdot IR \cdot EF \cdot ED}{BW \cdot AT}$$

C = Concentração do contaminante

- Quantos dados e como coletar dados para ter uma estimação fidedigna do valor de C?
  - Quantos dados - Suficiência
  - Estimação fidedigna - Representatividade

EPA United States Environmental Protection Agency

Comparação e apresentação de diferentes Designs de Amostragem
(viabilidade, vantagens, limitações, objetivos de cada um)

Guidance on Choosing a Sampling Design for Environmental Data Collection for use in Developing a Quality Assurance Project Plan

Mostra **como escolher um Design** em função do problema avaliado

ESTUDOS DE RISCO À SAÚDE HUMANA

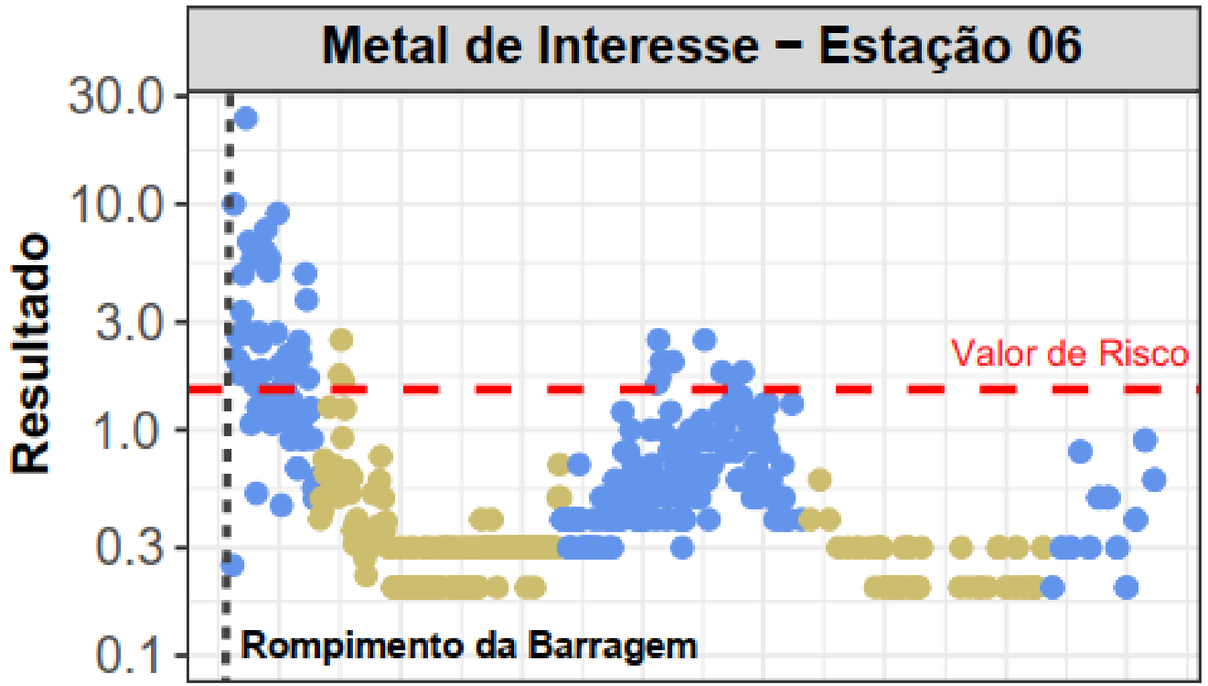
Dados coletados e/ou dados secundários já “em mãos”

Como saber se os dados que tenho estão suficientes e são representativos para os estudos de risco?

Mais de uma fonte de dados

- Comparabilidade de dados
- Temporalidade de dados
- Espacialidade das concentrações
- Avaliação de níveis basais (background) e conhecimento prévio (literatura ou avaliação de resultados em regiões “não-expostas” ou “pré-evento”)
- Inferências para comparações com valores de risco, unidades de exposição e áreas
- Avaliações de valores extremos

Temporalidade

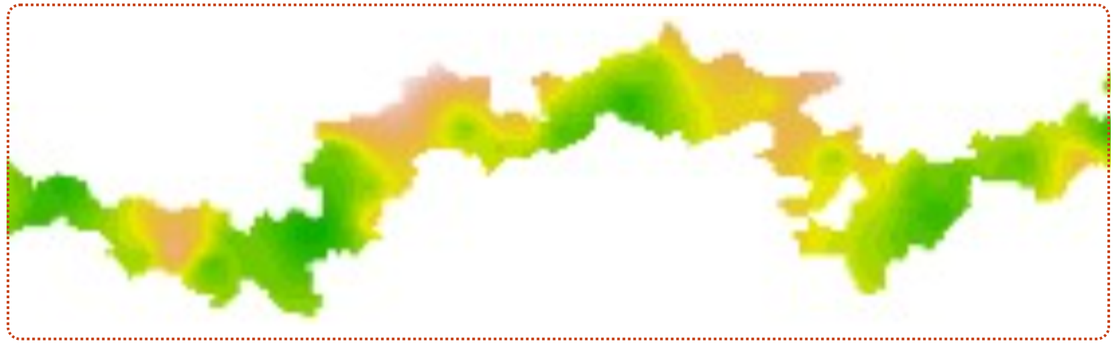


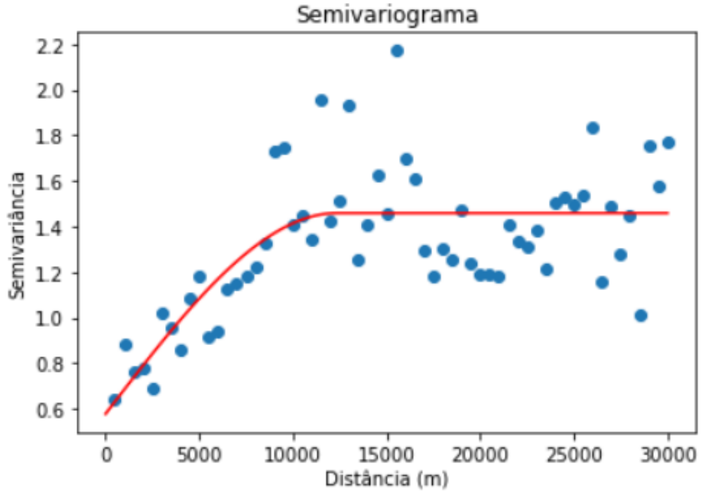
ESTUDOS DE RISCO À SAÚDE HUMANA

Dados coletados e/ou dados secundários já “em mãos”

Como saber se os dados que tenho estão suficientes e são representativos para os estudos de risco?

- Comparabilidade de dados
- Temporalidade de dados
- **Espacialidade das concentrações**
- Avaliação de níveis basais (background) e conhecimento prévio (literatura ou avaliação de resultados em regiões “não-expostas” ou “pré-evento”)
- Inferências para comparações com valores de risco, unidades de exposição e áreas
- Avaliações de valores extremos



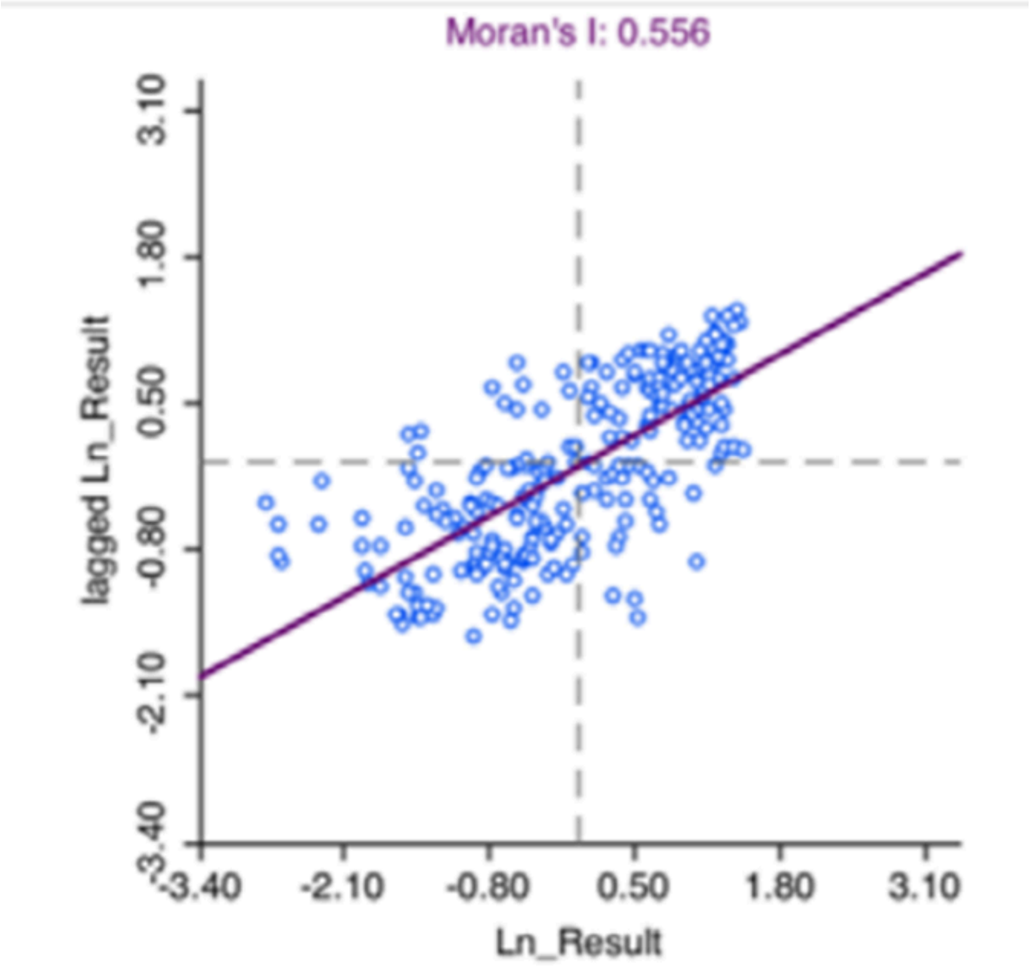


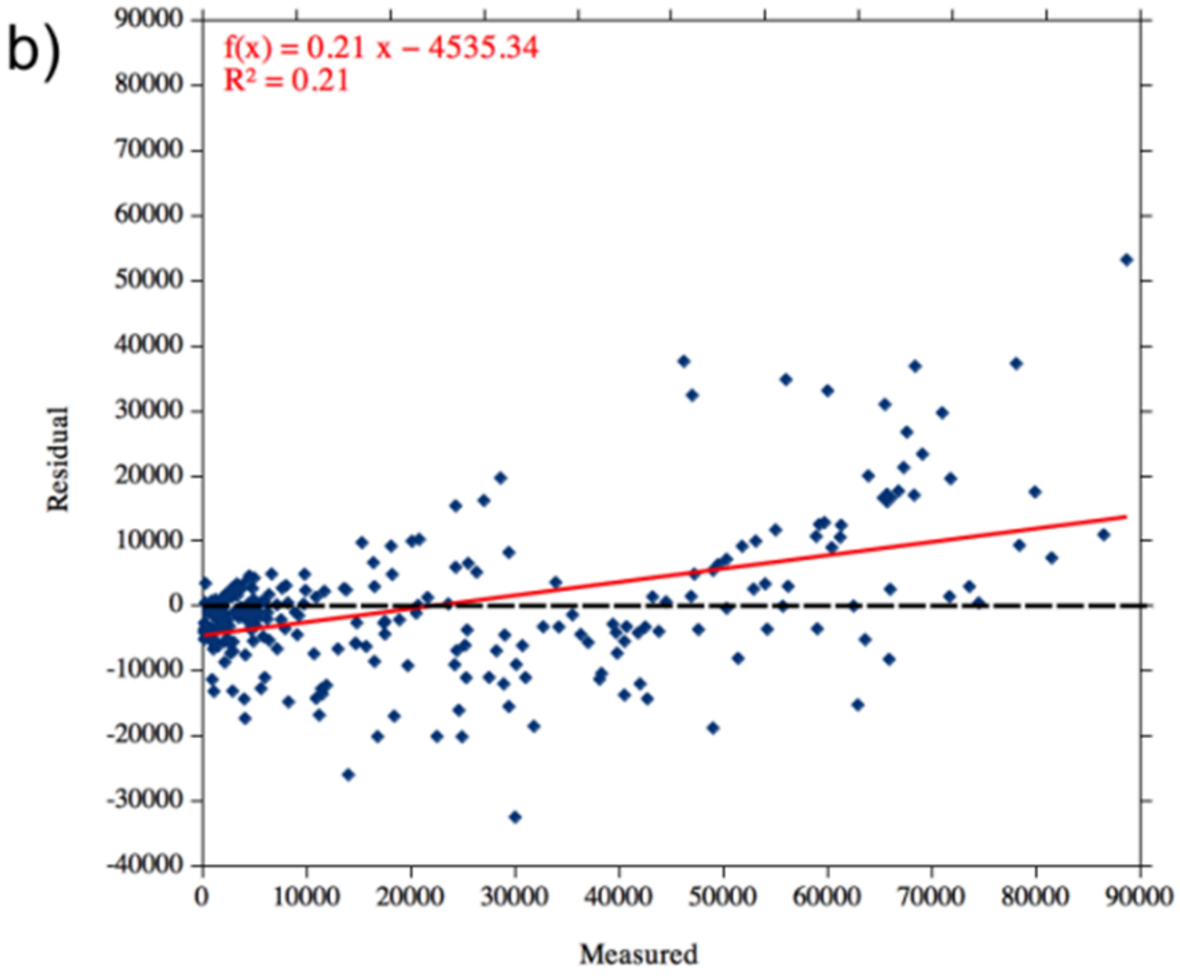
Distribuição dos dados

Análise variográfica

Resíduos, validação cruzada

ESTUDOS DE RISCO À SAÚDE HUMANA



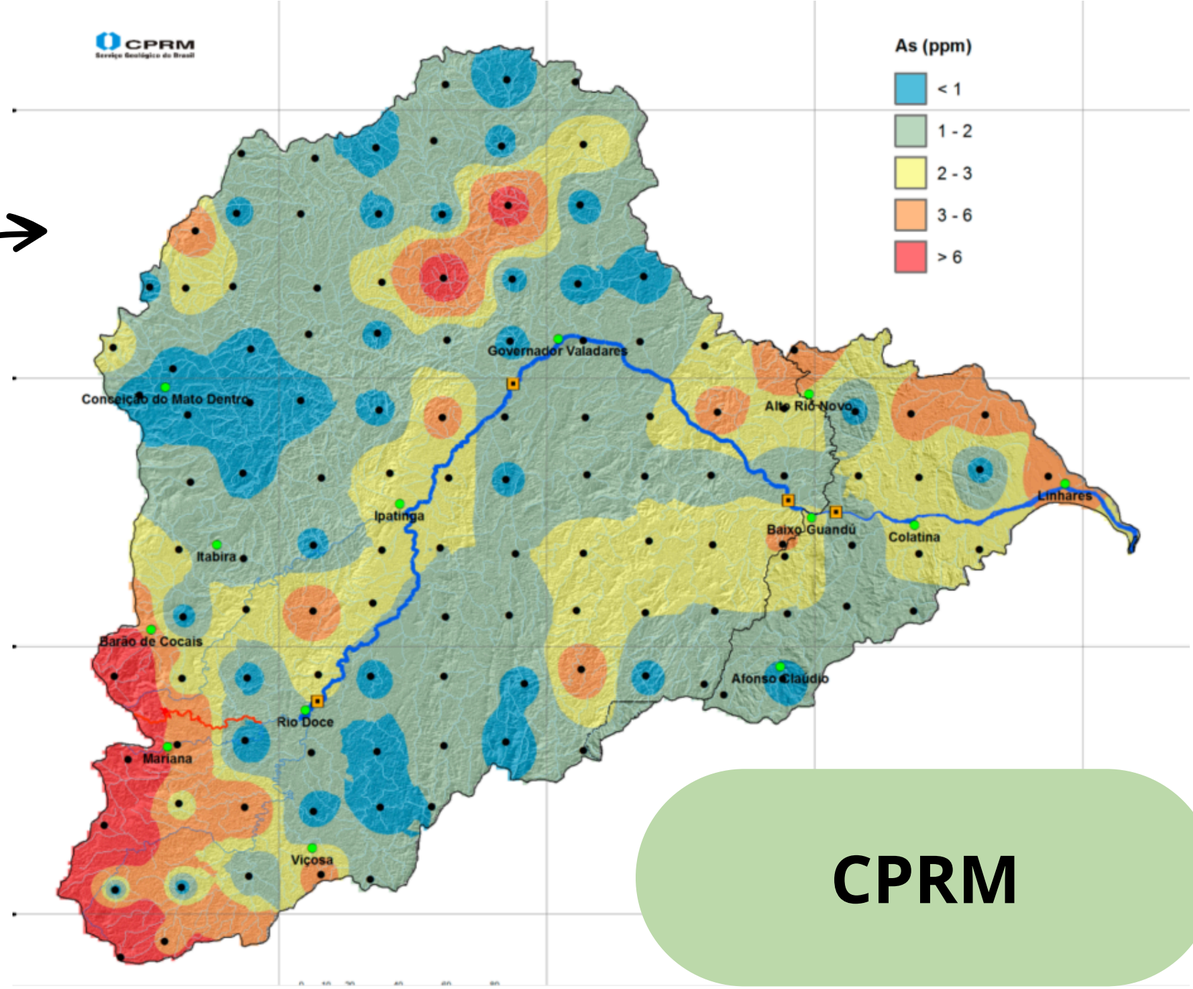


ESTUDOS DE RISCO À SAÚDE HUMANA

Dados coletados e/ou dados secundários já “em mãos”

Como saber se os dados que tenho estão suficientes e são representativos para os estudos de risco?

- Comparabilidade de dados
- Temporalidade de dados
- Espacialidade das concentrações
- **Avaliação de níveis basais (background) e conhecimento prévio** (literatura ou avaliação de resultados em regiões “não-expostas” ou “pré-evento”)
- Inferências para comparações com valores de risco, unidades de exposição e áreas
- Avaliações de valores extremos



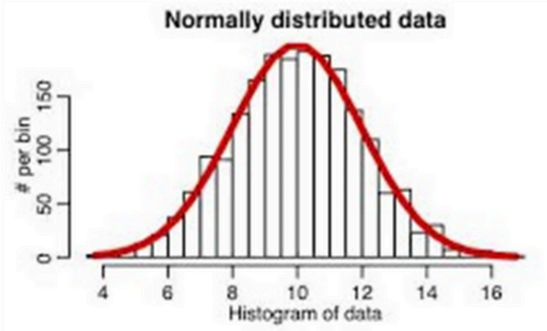
CPRM

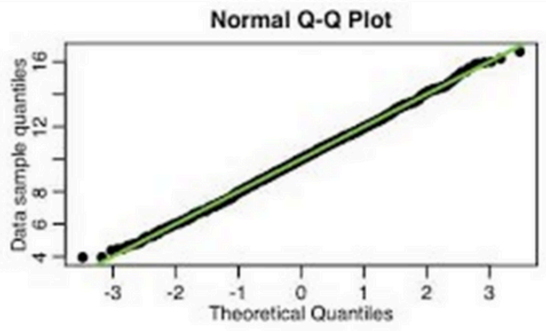
ESTUDOS DE RISCO À SAÚDE HUMANA

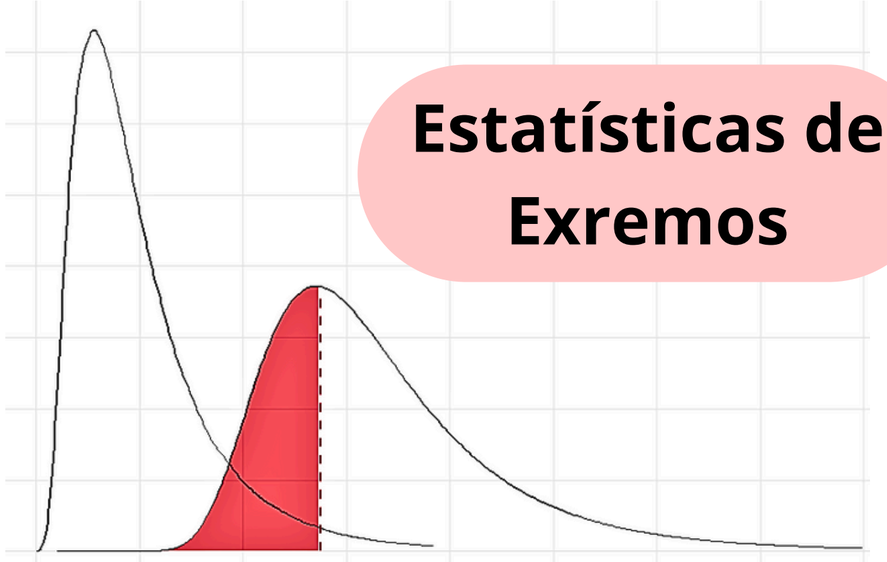
Dados coletados e/ou dados secundários já “em mãos”

Como saber se os dados que tenho estão suficientes e são representativos para os estudos de risco?

- Comparabilidade de dados
- Temporalidade de dados
- Espacialidade das concentrações
- Avaliação de níveis basais (background) e conhecimento prévio (literatura ou avaliação de resultados em regiões “não-expostas” ou “pré-evento”)
- **Inferências para comparações com valores de risco, unidades de exposição e áreas**
- **Avaliações de valores extremos**





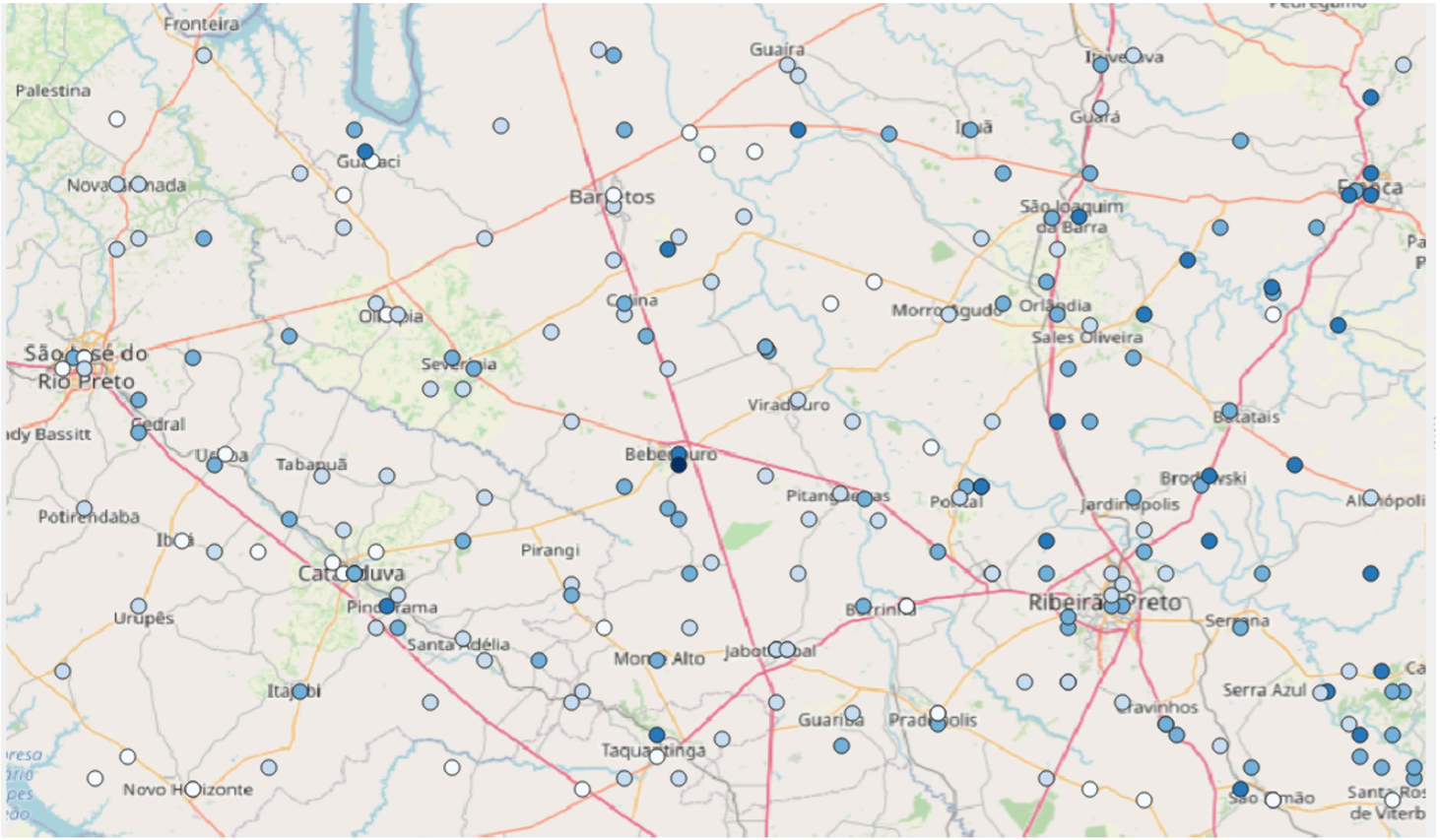


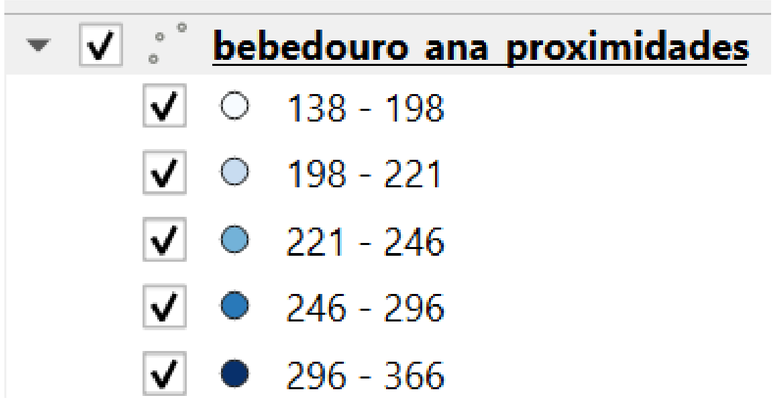
Estatísticas de Exremos

DADOS DE PRECIPITAÇÃO - ANA

POLÍGONOS DE THIESSEN OU DIAGRAMAS DE VORONOI

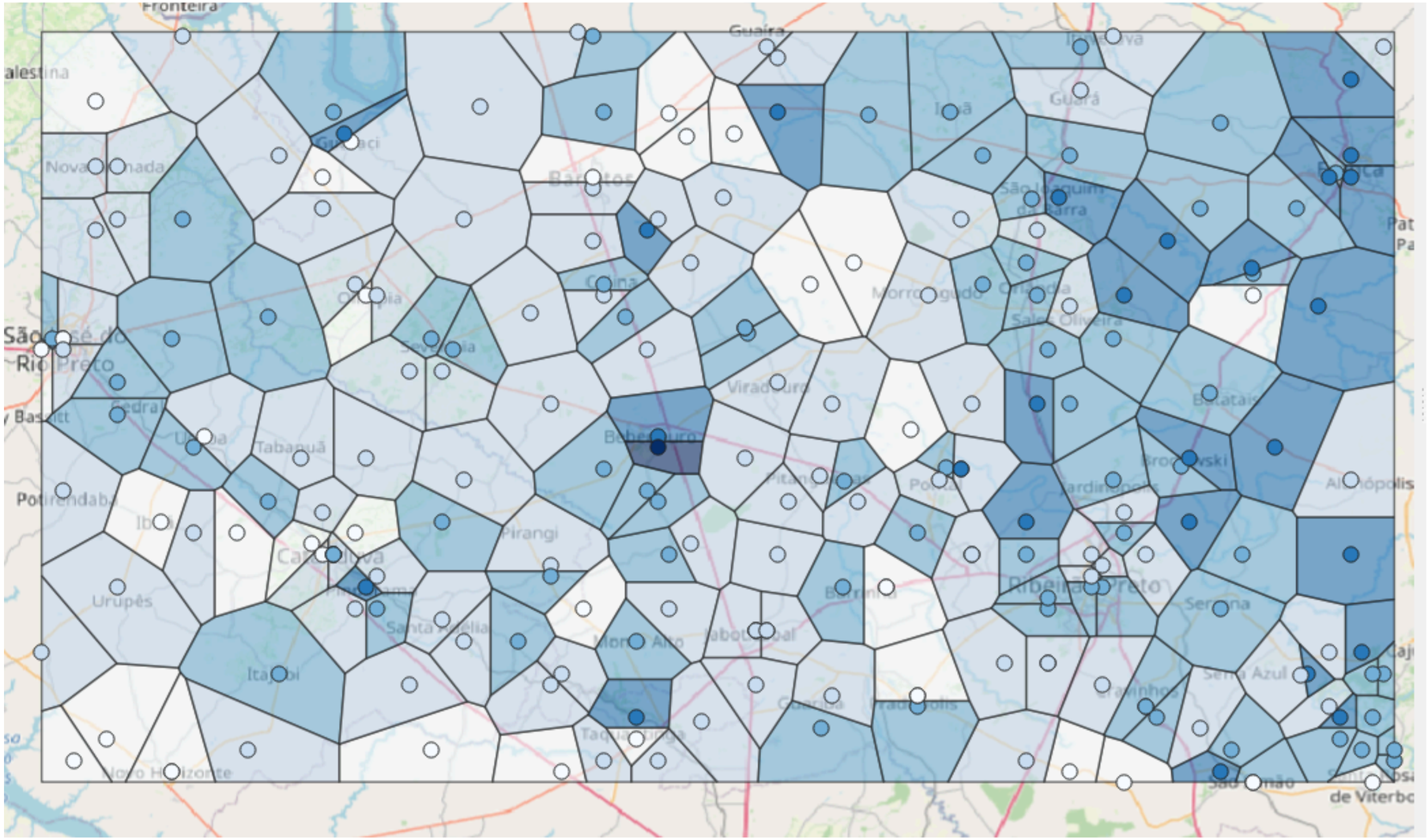
- Dados de **precipitação histórica** (recorte próximo à cidade de Bebedouro, interior de SP)
- Variável de interesse: Média de Fevereiro

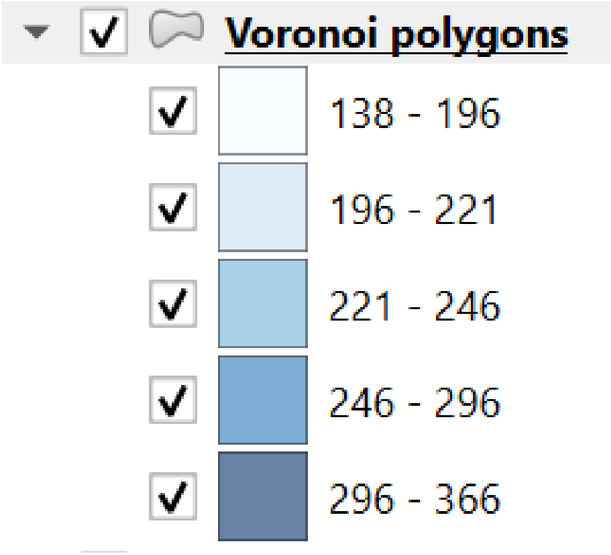




- Mapa com valores de precipitação média de fevereiro em pontos discretos

POLÍGONOS DE THIESSEN NO QGIS

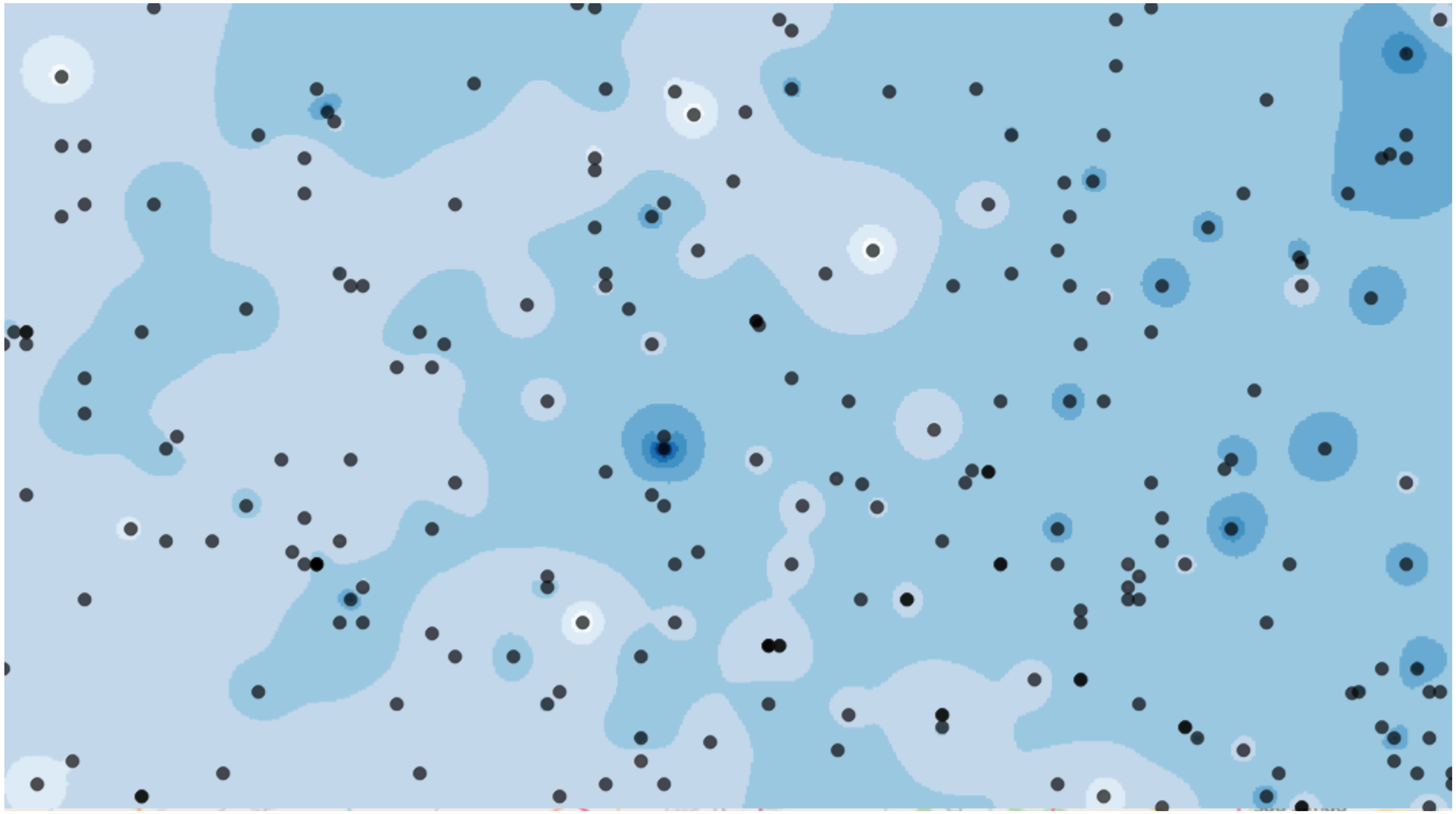




- Cada polígono assume que toda a chuva dentro dele é igual à da estação mais próxima

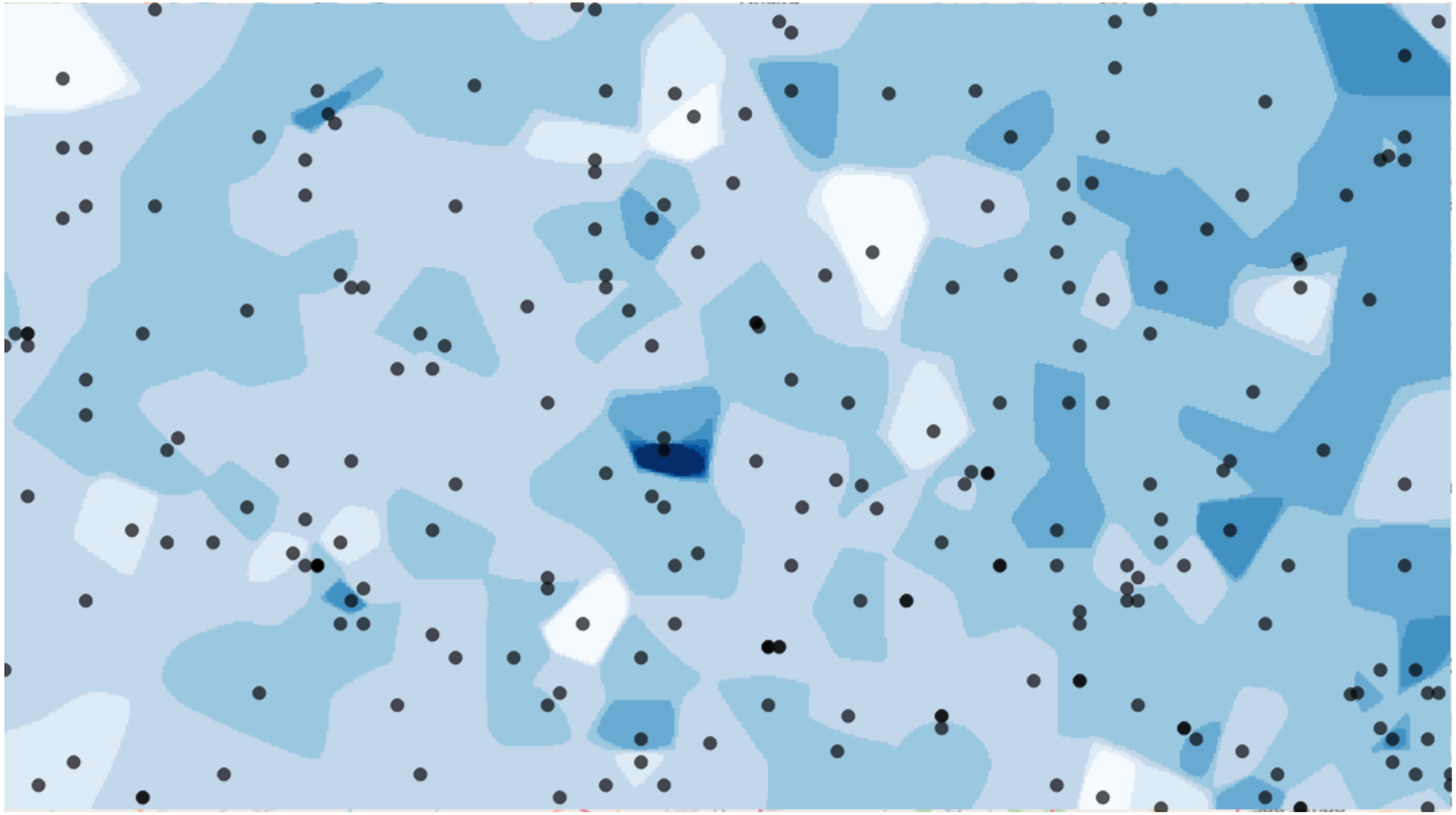
IDW NO QGIS

IDW (Potência 2)



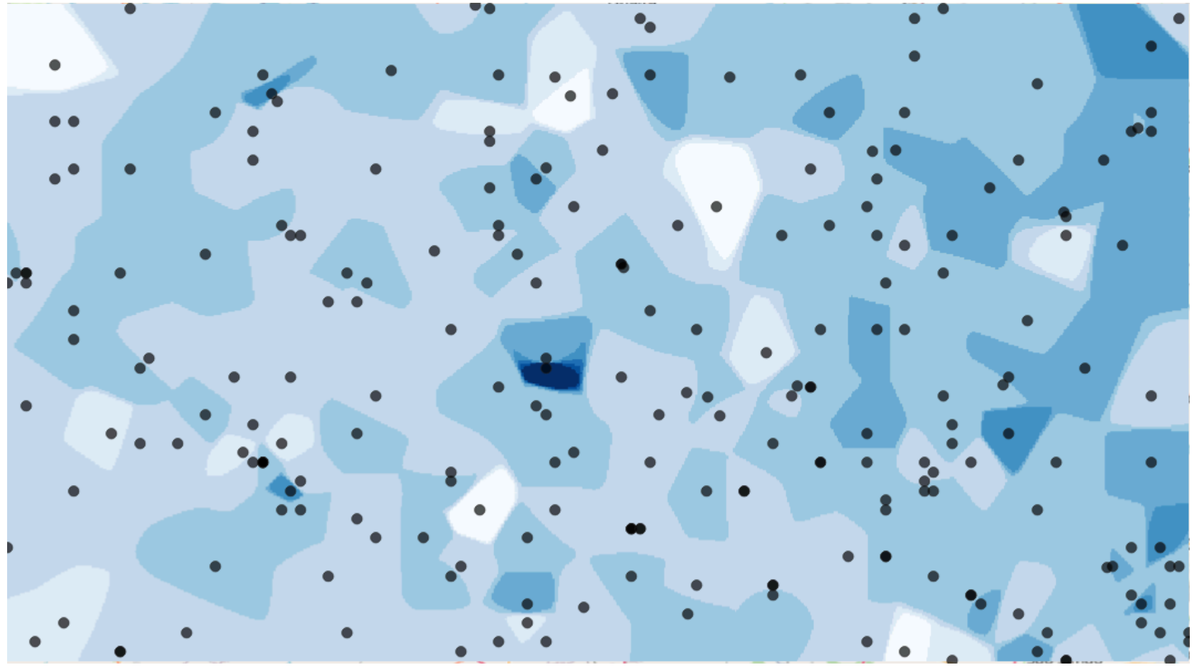
IDW NO QGIS

IDW (Potência 15)



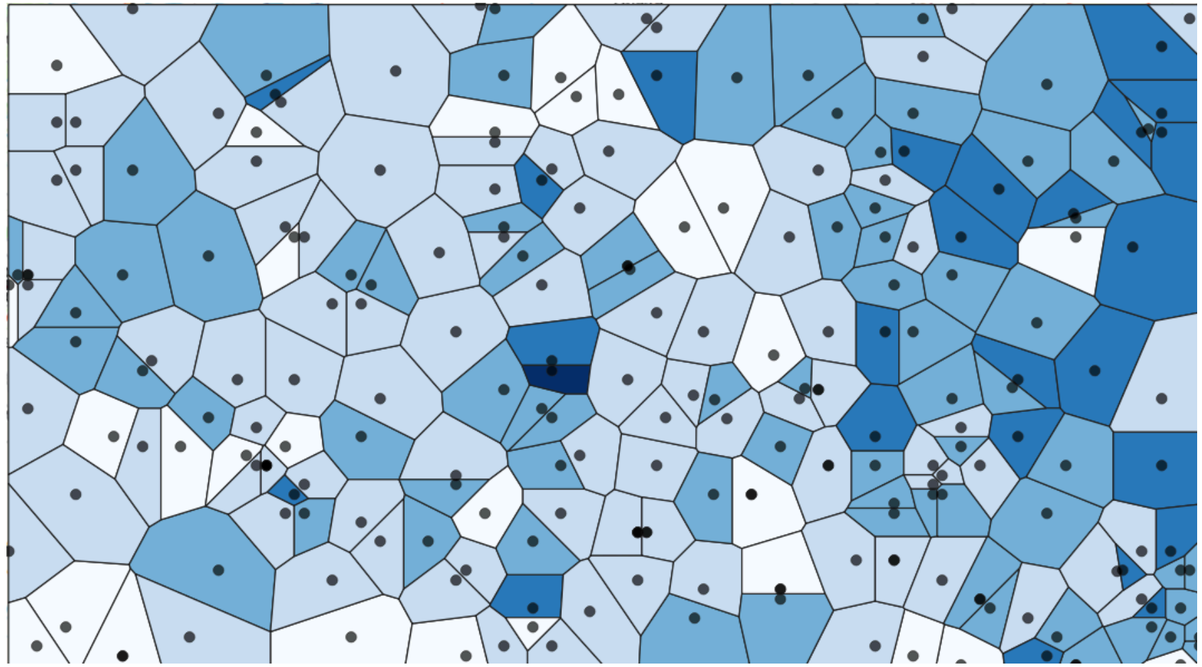
IDW NO QGIS

IDW (Potência 15)



- Método **suave**, mas com forte influência de pontos próximos
- **Valores caem rapidamente** conforme a distância aumenta
- Resultados muito próximos com aumento do expoente p
- Ainda há interpolação, mas a superfície fica mais segmentada, semelhante aos Thiessen

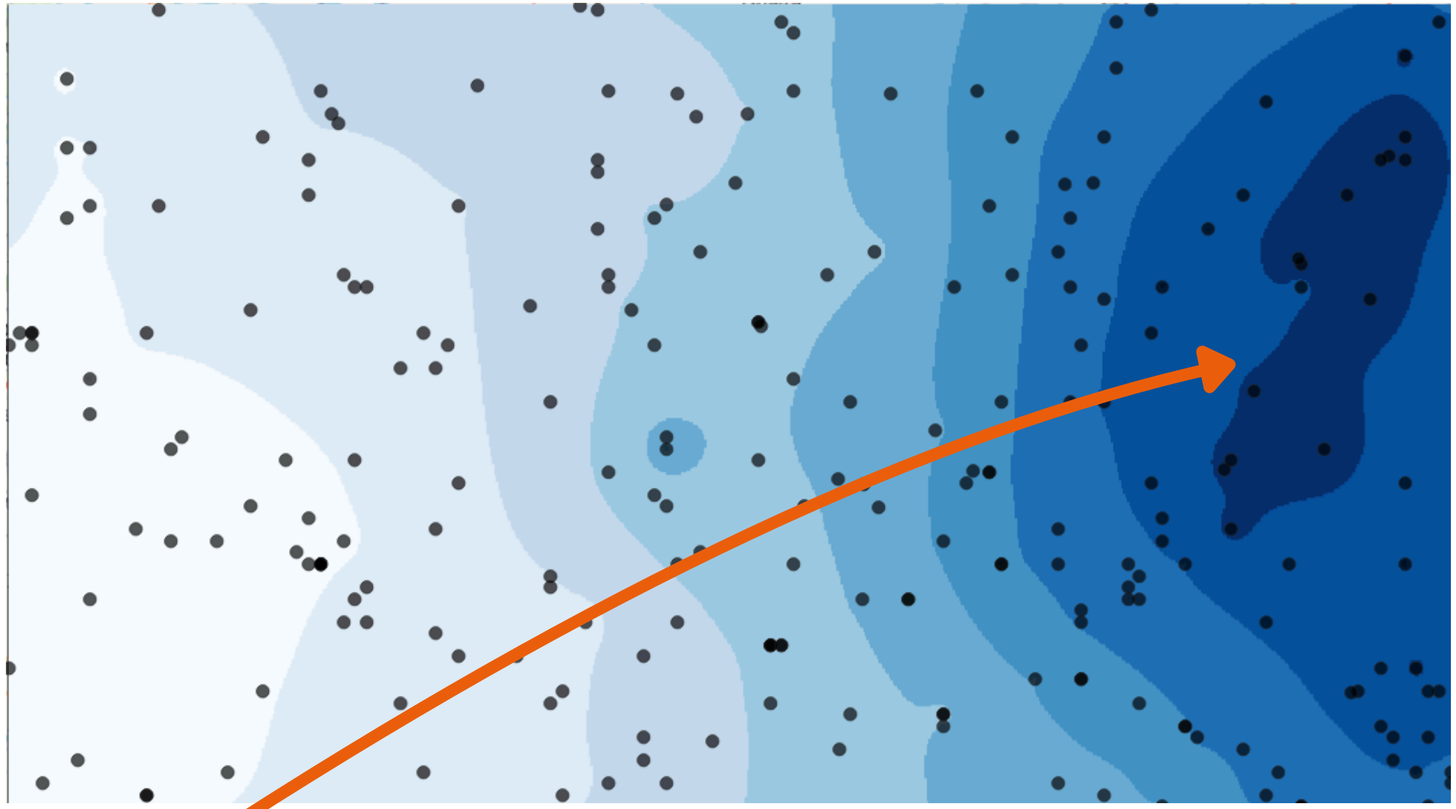
Polígonos de Thiessen



- Método **abrupto** - cada ponto assume o valor mais próximo
- **Não há transição gradual** entre os valores
- Criam **fronteiras rígidas**, sem variações dentro de cada polígono
- Melhor para zonas de influência absolutas (ex: qual estação pluviométrica cobre cada região)

IDW NO QGIS

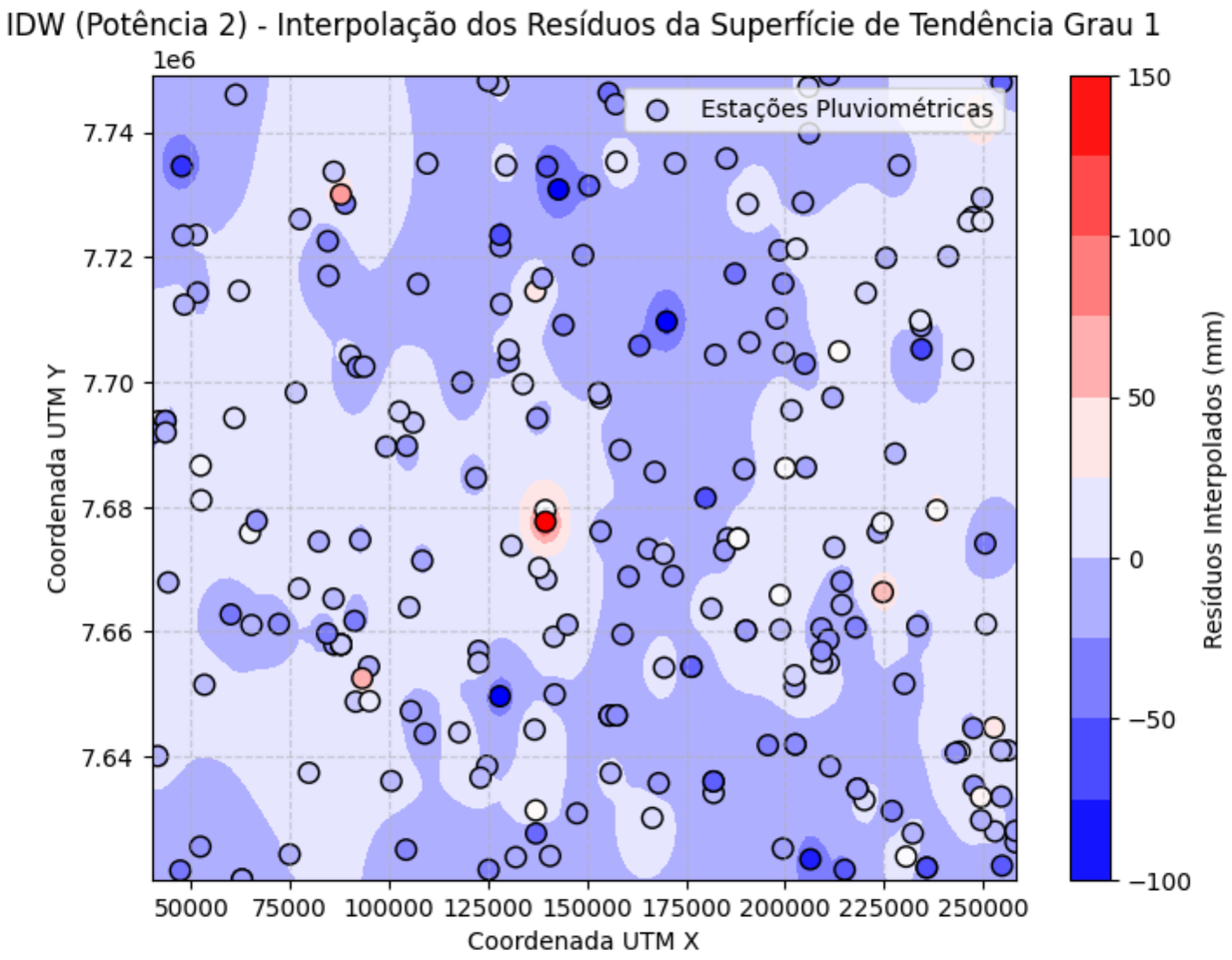
IDW (Potência 0,01) → **Extremamente suavizada**

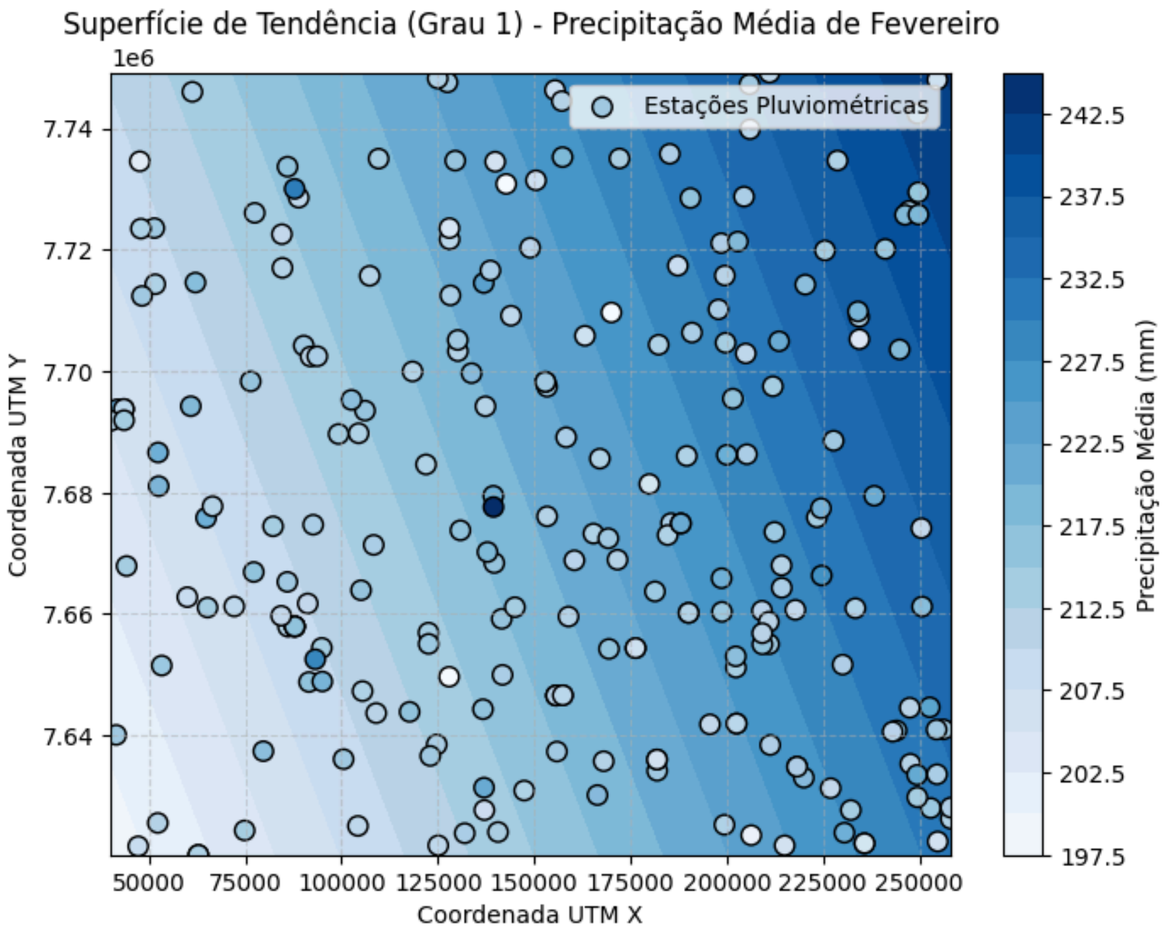


Direção espacial: Gradiente de precipitação, com valores mais altos concentrados em certas regiões (tons mais escuros de azul) e valores mais baixos em outras (tons mais claros).

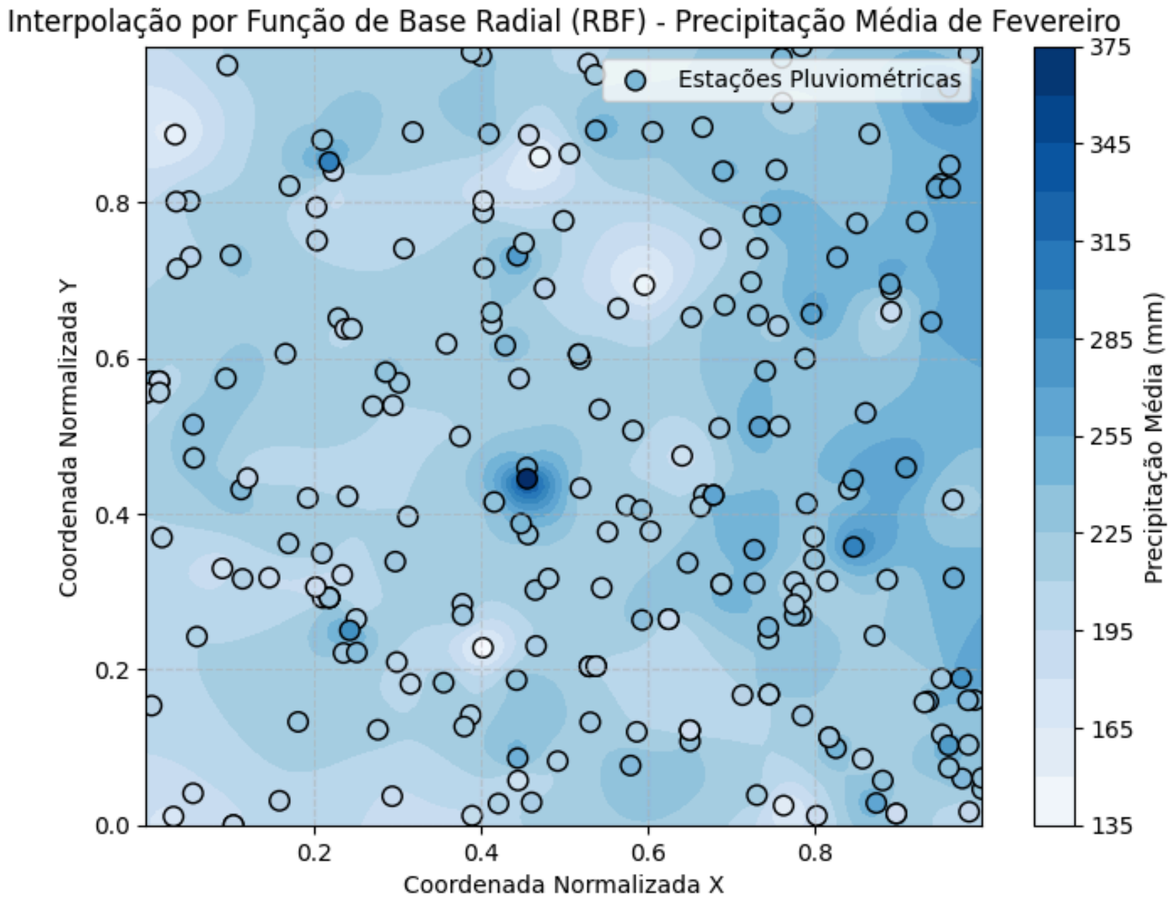
- **Suavização extrema**: Como expoente p é muito pequeno, a influência dos pontos mais distantes é quase igual à dos mais próximos
- Isso gera uma interpolação em que os valores se misturam de forma gradual, sem focar em regiões muito específicas.
- **Padrão mais contínuo** - Contornos bem suaves e não há grandes variações abruptas
- **Menos peso para outliers** - Caso existam, pontos isolados tem pouca influência individual (pois a interpolação distribui os pesos de forma ampla)
- A transição suave indica uma **possível influência de padrões regionais**

SUPERFÍCIES DE TENDÊNCIA NO PYTHON

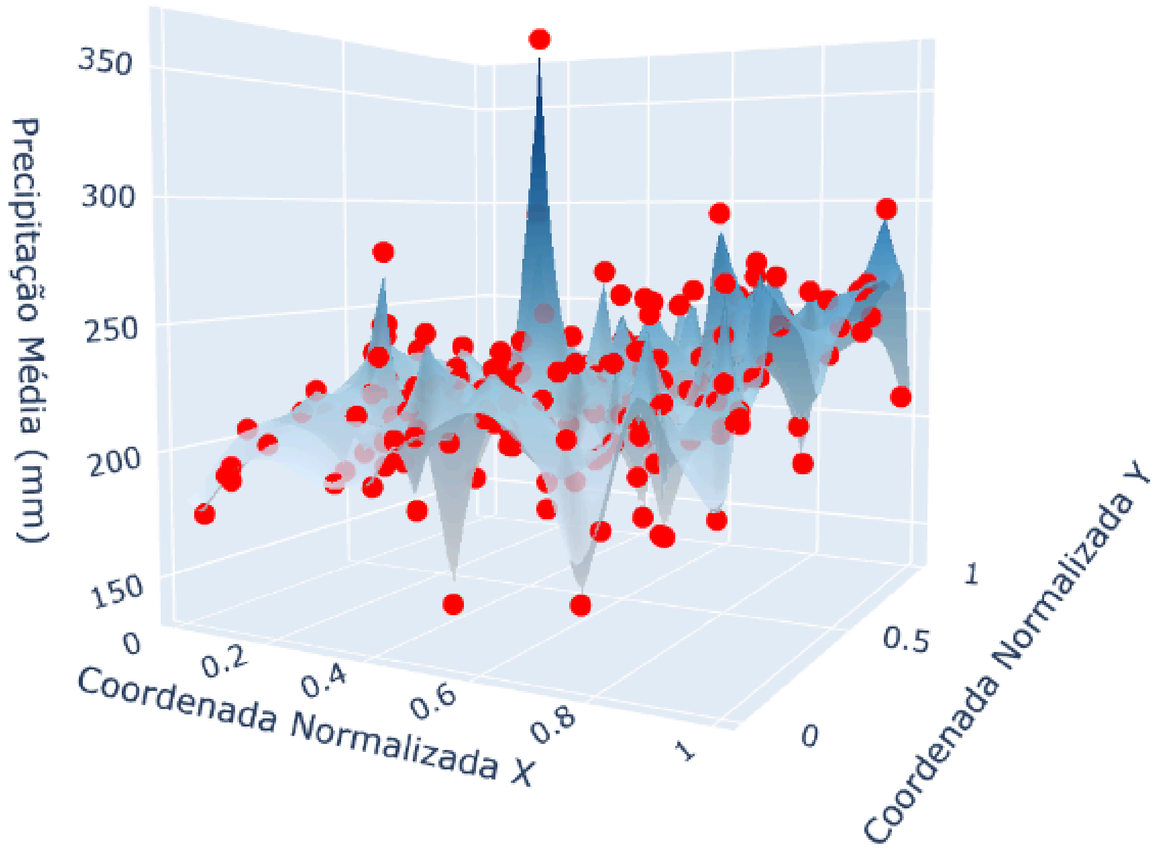


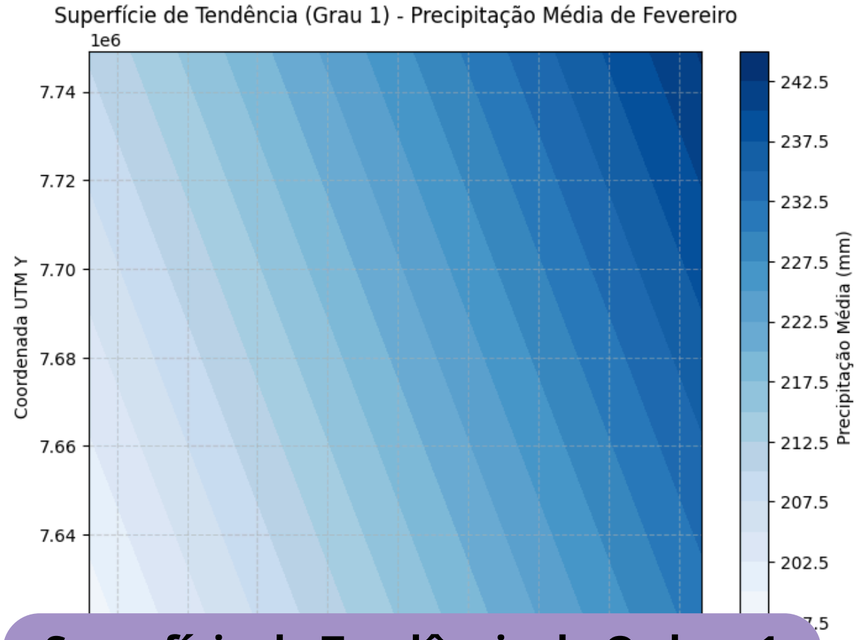


FUNÇÕES DE BASE RADIAL (RBF)

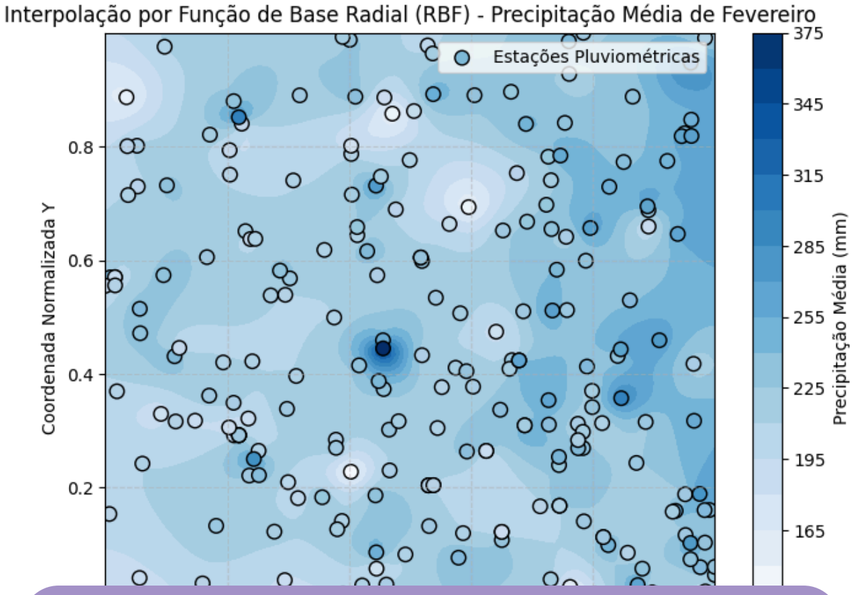


RBF Linear

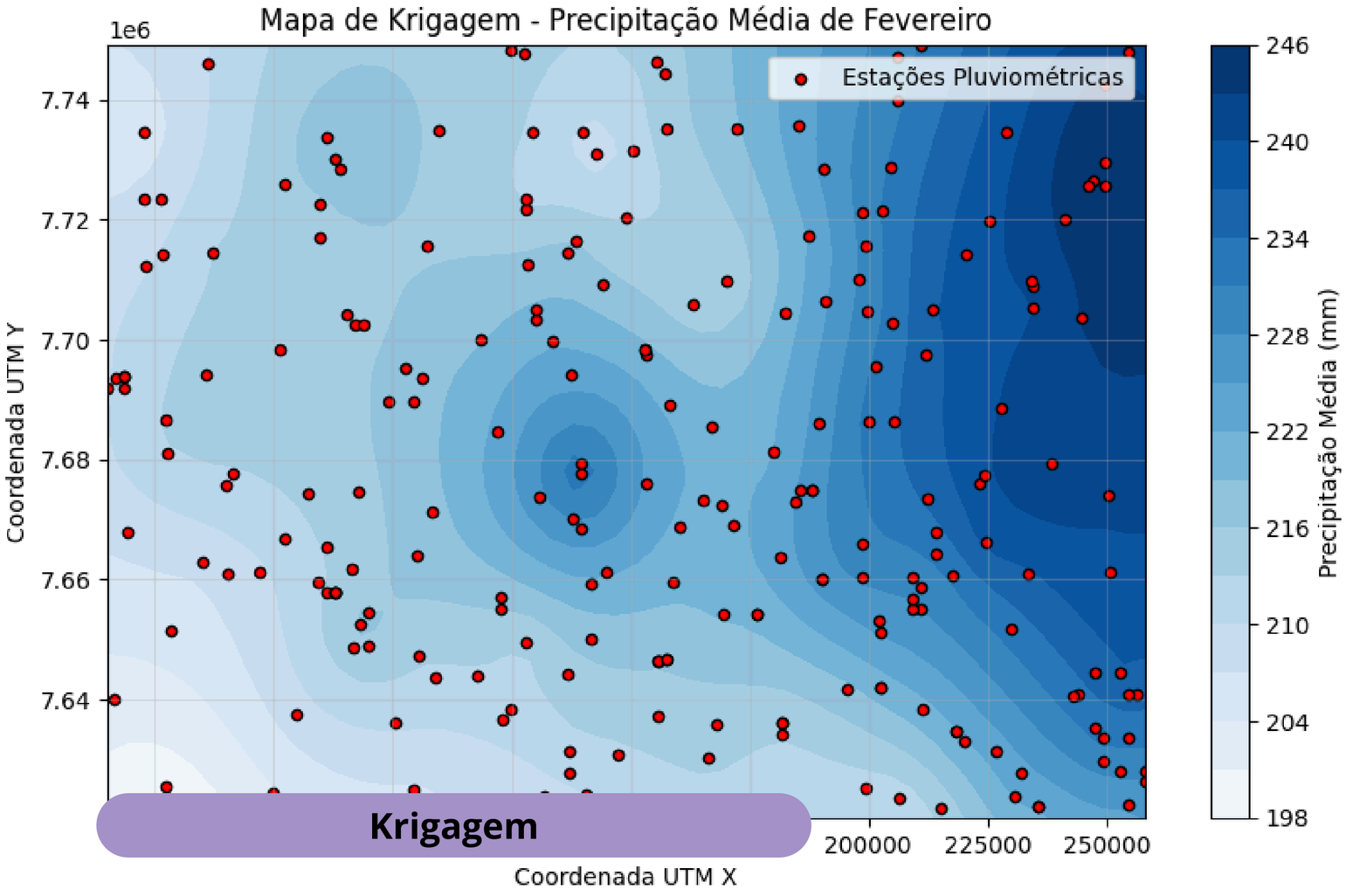




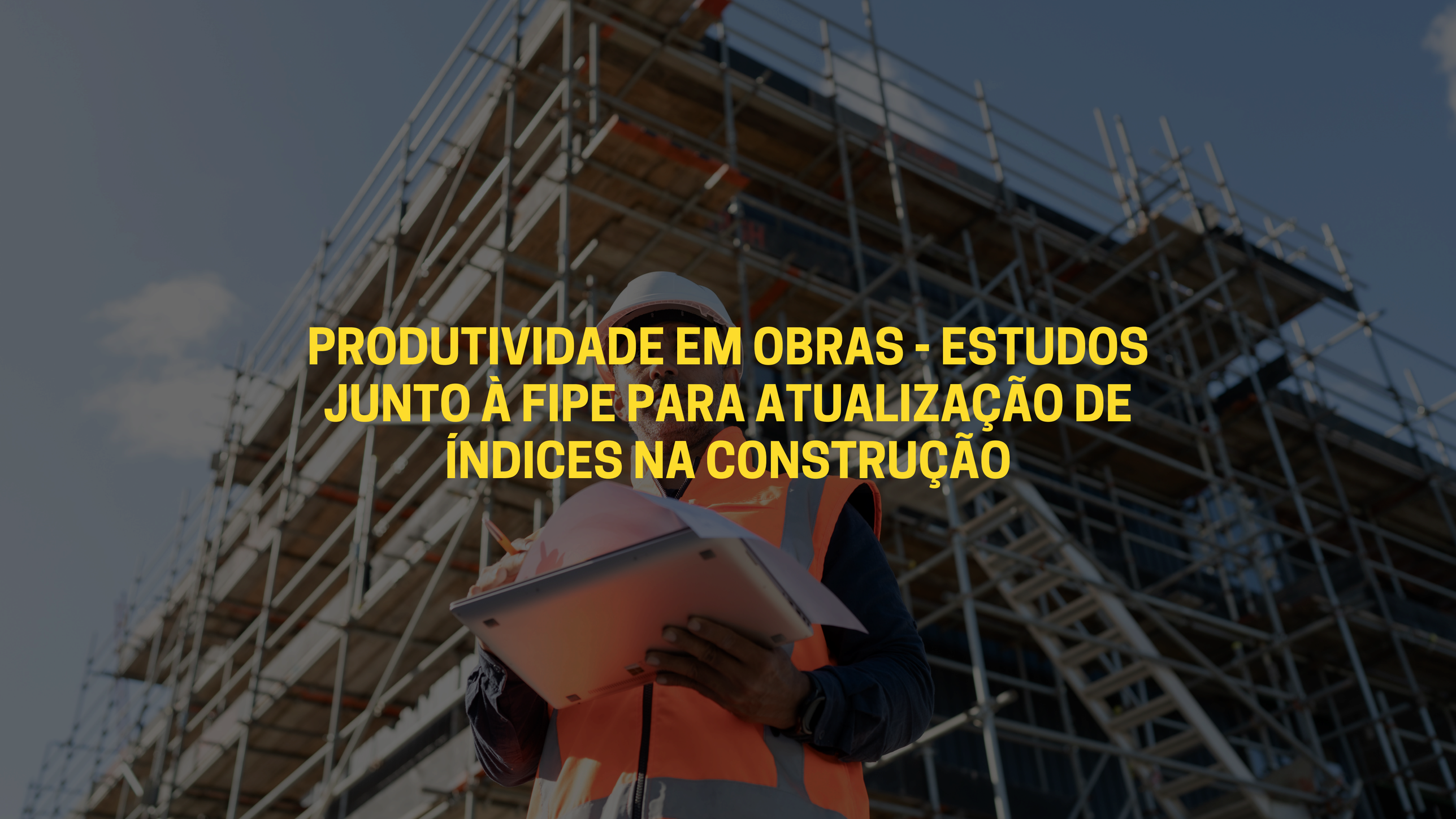
Superfície de Tendência de Ordem 1



Função de Base Radial (RBF)



Krigagem



PRODUTIVIDADE EM OBRAS - ESTUDOS JUNTO À FIPE PARA ATUALIZAÇÃO DE ÍNDICES NA CONSTRUÇÃO

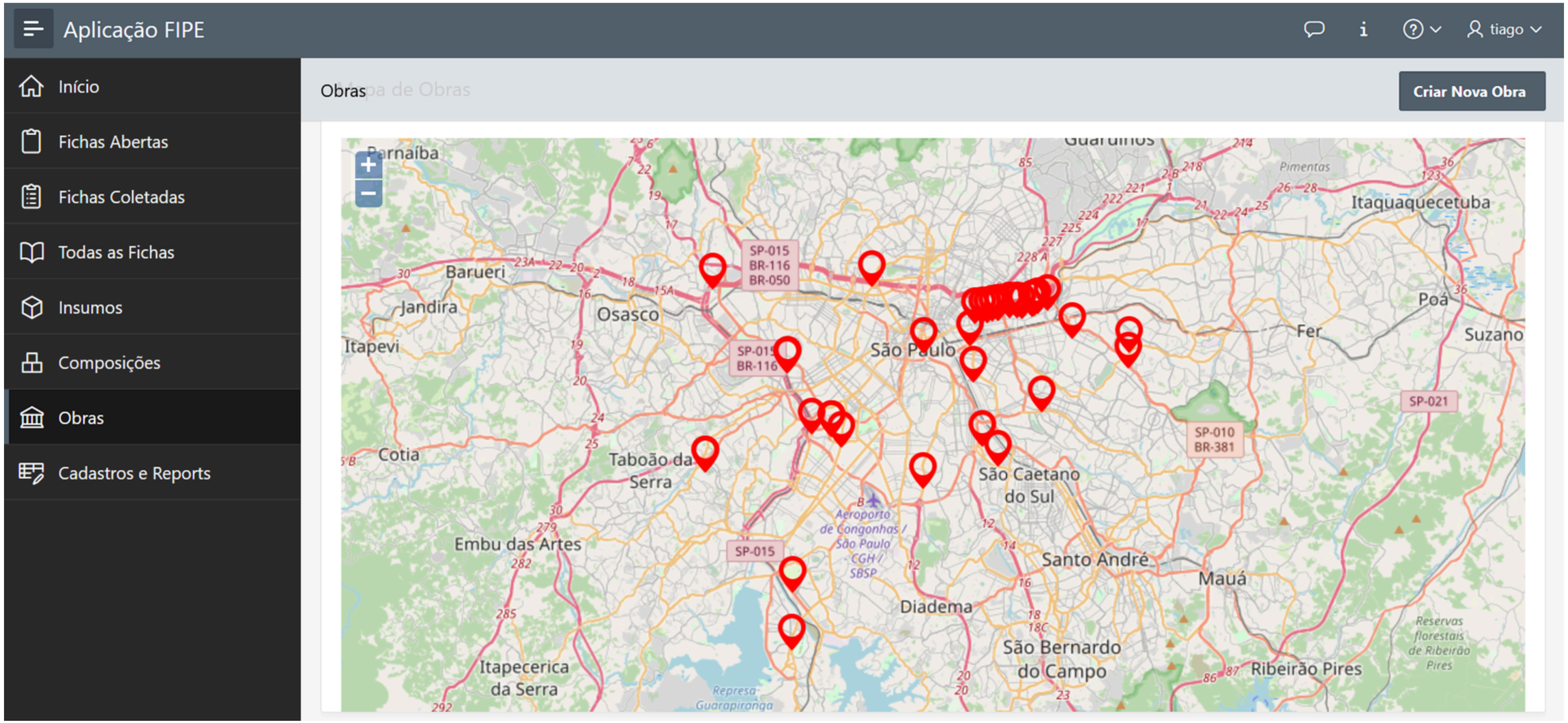
ENGENHARIA CIVIL - PRODUTIVIDADE EM OBRAS

Qual é a produtividade dos serviços de construção civil em São Paulo? Que valor devemos colocar na tabela de orçamento?

Para responder essas perguntas, tivemos que entender primeiro:

- Quais são os principais serviços da construção?
- Quais são aqueles que custam mais?
- Como coletar dados de produtividade? Quantos dados devo coletar? Uma obra é suficiente, mais de uma, várias?
- E pra serviços que são altamente complexos, é o mesmo cálculo?

ENGENHARIA CIVIL - PRODUTIVIDADE EM OBRAS



ENGENHARIA CIVIL - PRODUTIVIDADE EM OBRAS

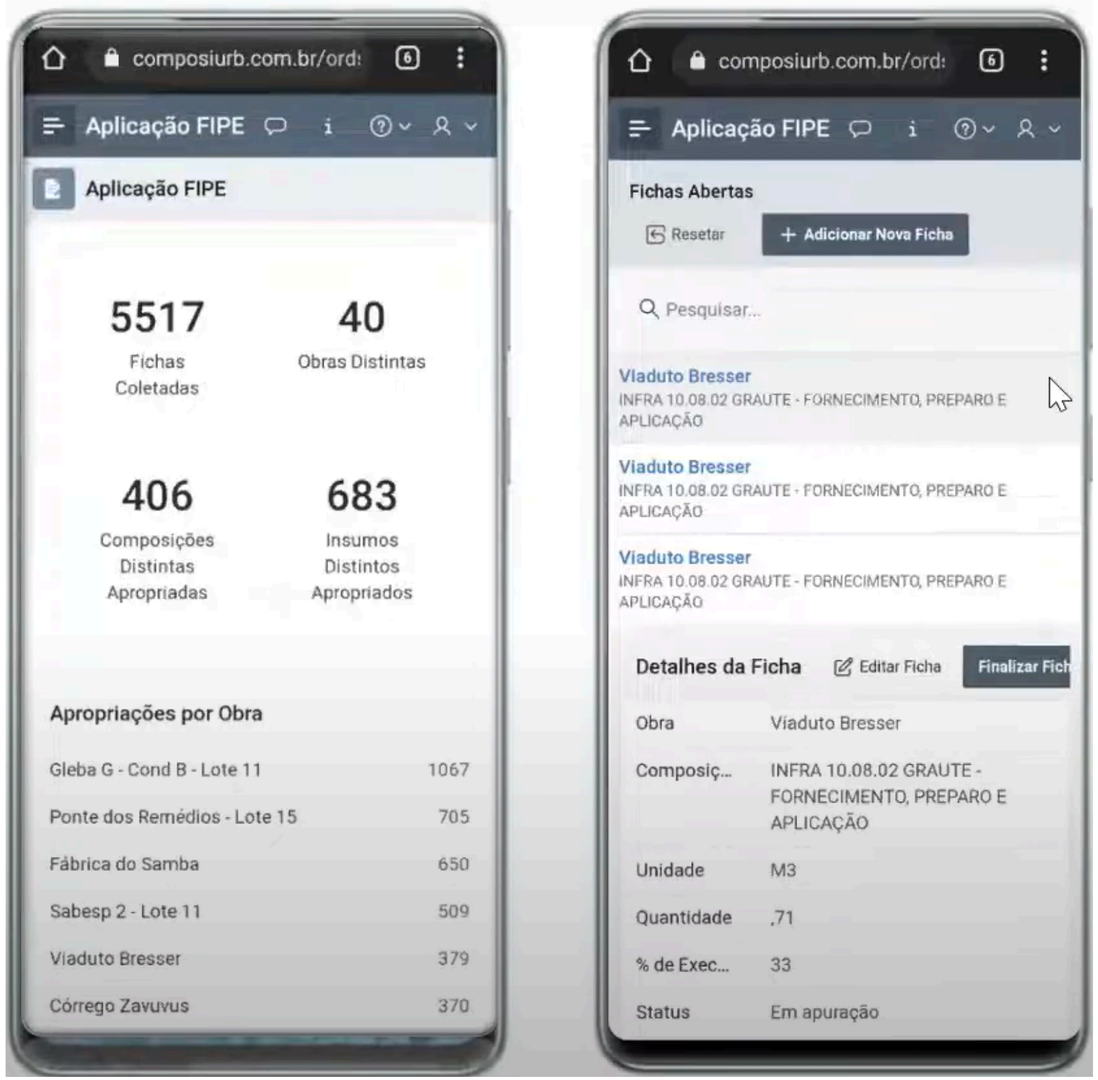
Coleta de **mais de 800h de atividades em obras** em várias obras diferentes

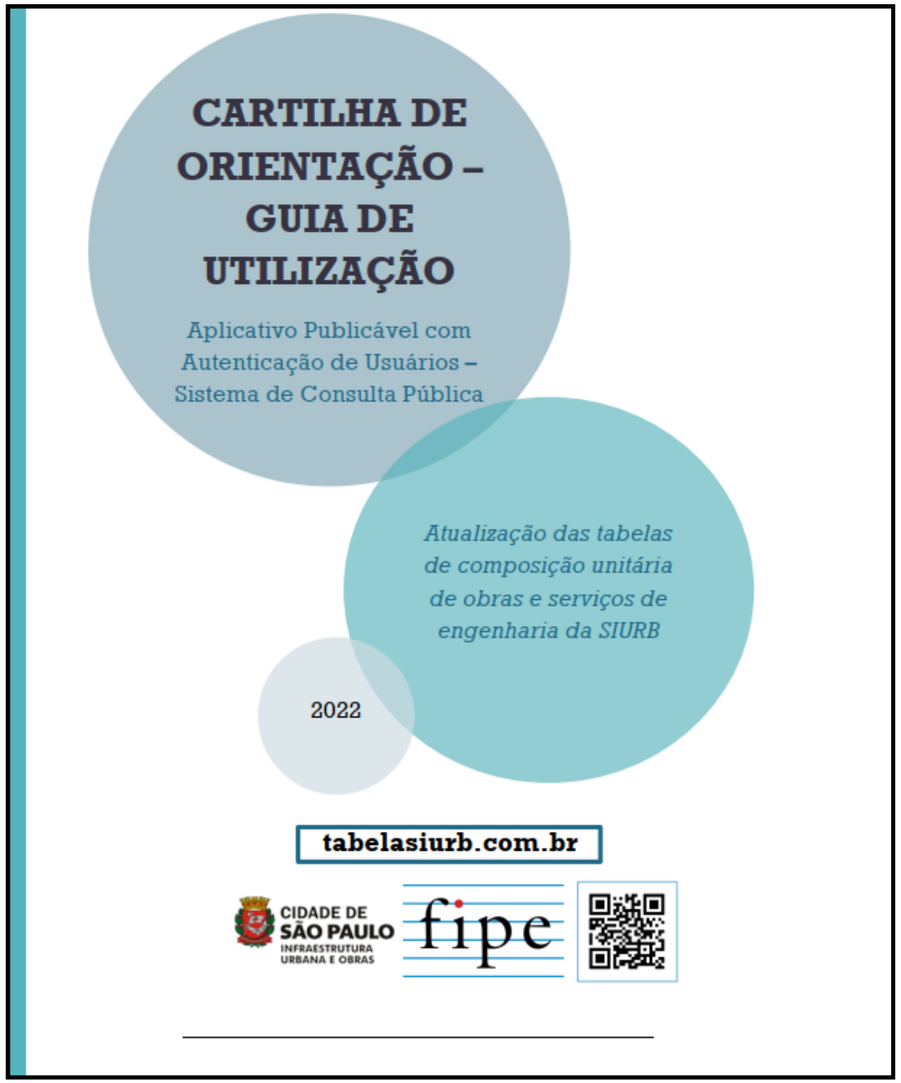
Exemplo de serviço: Alvenaria

Acompanhamento em obras, cálculos com os valores obtidos, comparação e ponderação com referências científicas e de outros lugares, obtenção do valor a ser adotado na cidade

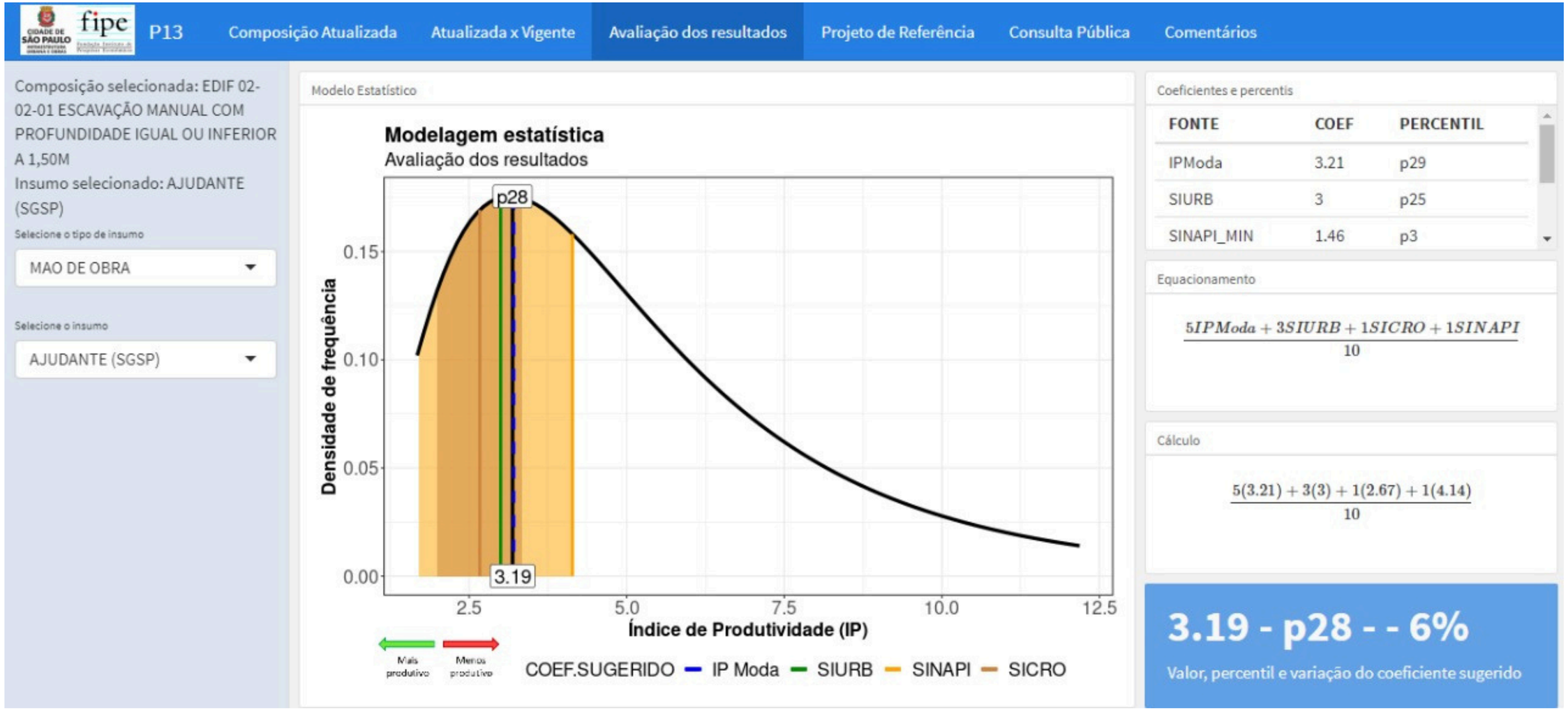
CÓDIGO	DESCRIÇÃO	UNID	COEF SIURB	COEF SUGERIDO	Δ
2020	PEDREIRO (SGSP)	H	0,80	0,48	+40%
2099	SERVENTE (SGSP)	H	0,80	0,46	+43%
-	ARGAMASSA ESTRUTURAL EM SILO	KG	-	*	*
12601	BLOCO DE CONCRETO ESTRUTURAL DE 14CM - RESISTÊNCIA 8MPA	Un	13,10	*	*

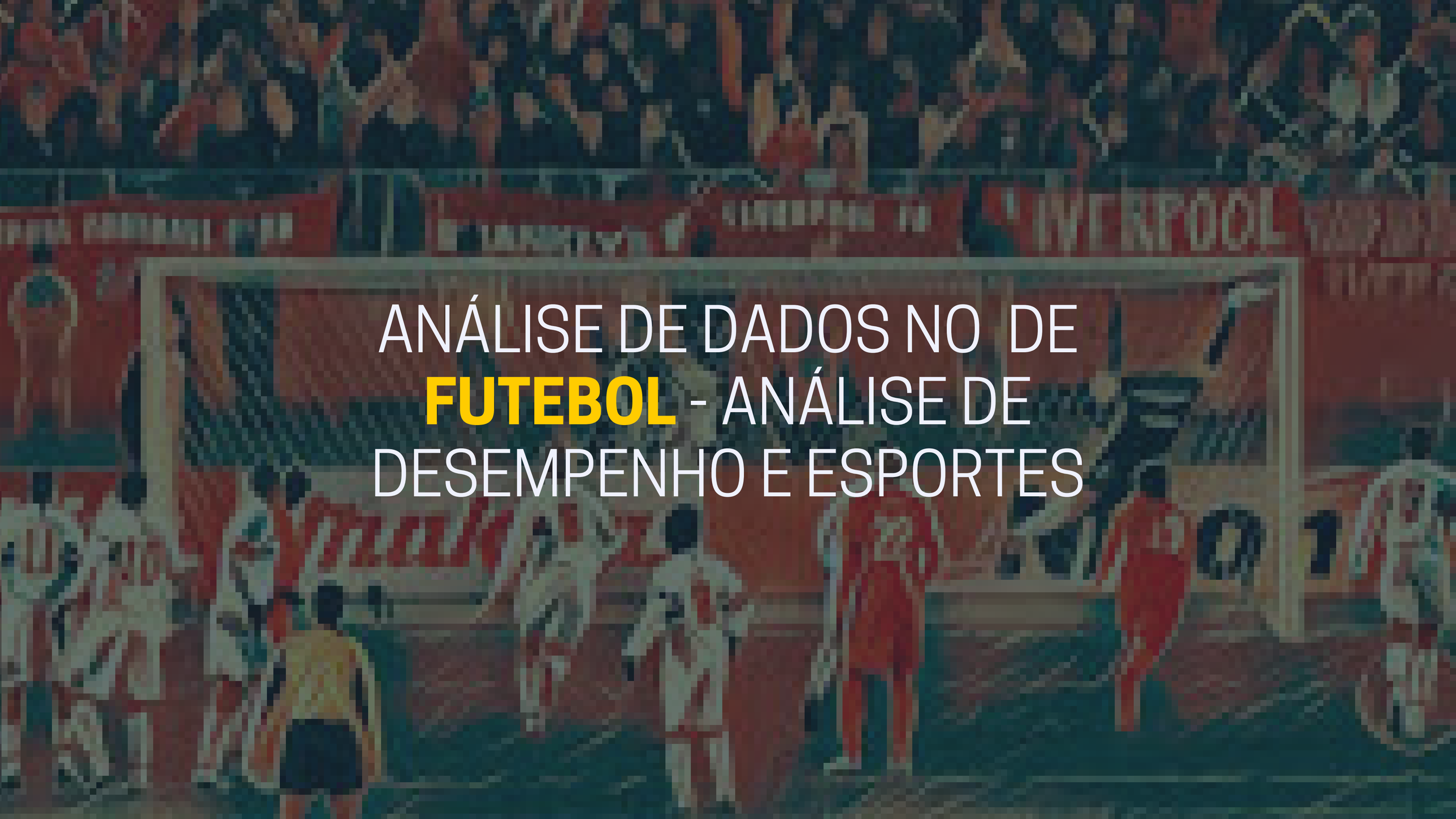
ENGENHARIA CIVIL - PRODUTIVIDADE EM OBRAS





ENGENHARIA CIVIL - PRODUTIVIDADE EM OBRAS





ANÁLISE DE DADOS NO DE **FUTEBOL** - ANÁLISE DE DESEMPENHO E ESPORTES

VARIÁVEIS DISPONÍVEIS NO BANCO DE DADOS

Variáveis Coletadas e Resumidas após tratamento

Um resultado por jogo de cada time

Sobre a partida

Time alvo, adversário, temporada (início/meio/final), mando de campo (casa/fora), estádio, campeonato, categoria campeonato (copa/estadual/pontos corridos/outros), mata-mata (sim/não), gols feitos, gols sofridos.

Sobre o desempenho

Defesas gk (quantidade), acurácia passes (mediana, %), acoes (quantidade), cruzamentos (quantidade), acurácia bolas longas (q3 %), divididas chão (mediana, %), divididas ar (mediana, %), faltas cometidas (quantidade), afastadas de bola (quantidade), faltas sofridas (quantidade), dribles (quantidade), chutes (quantidade), pontaria (quantidade)

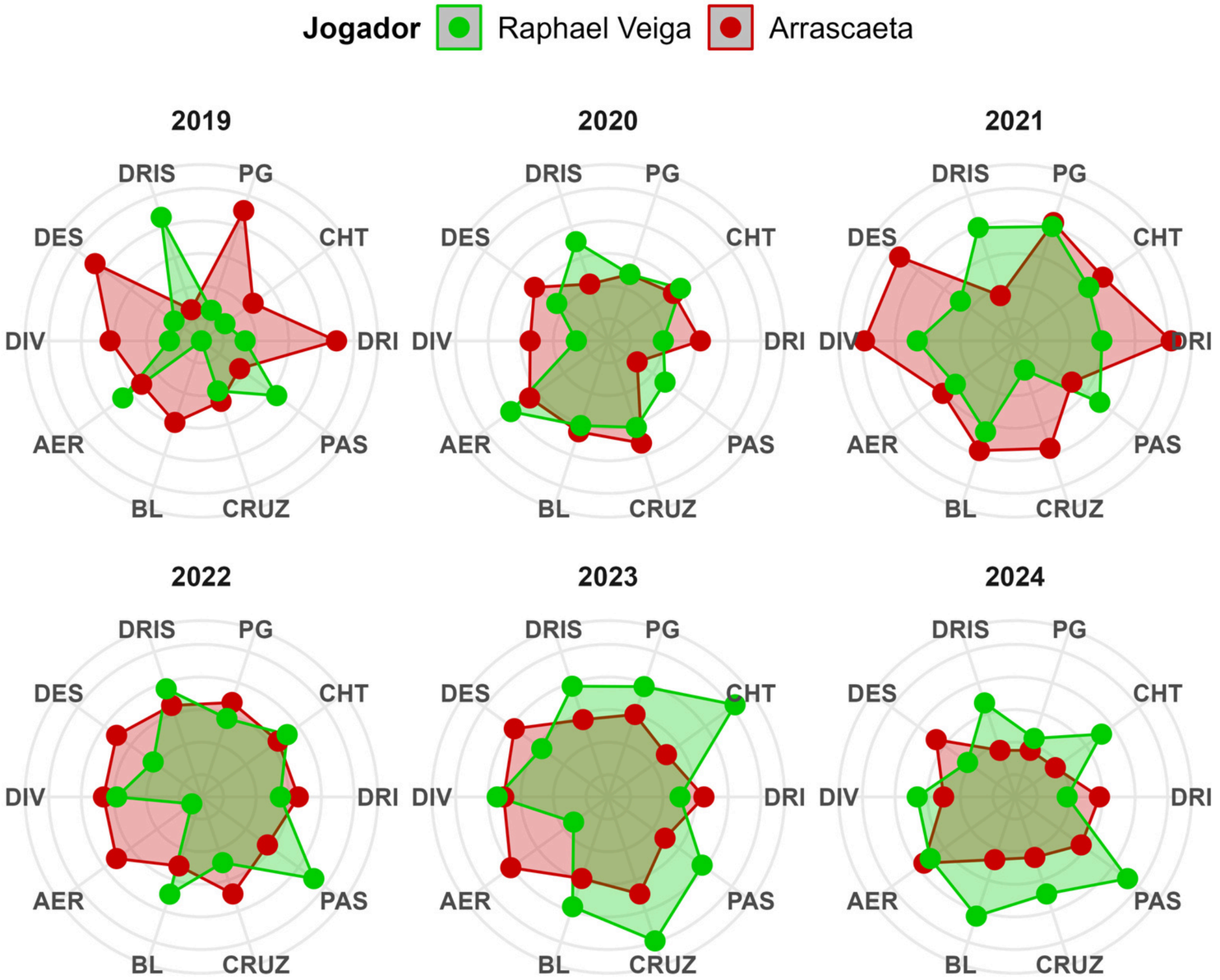
Time alvo: Ituano

PAULO MASSINI E THM ESTATÍSTICA

VEIGA X ARRASCAETA

Raphael Veiga x Arrascaeta - De igual para igual

Gráfico de Radar - Atributos gerais



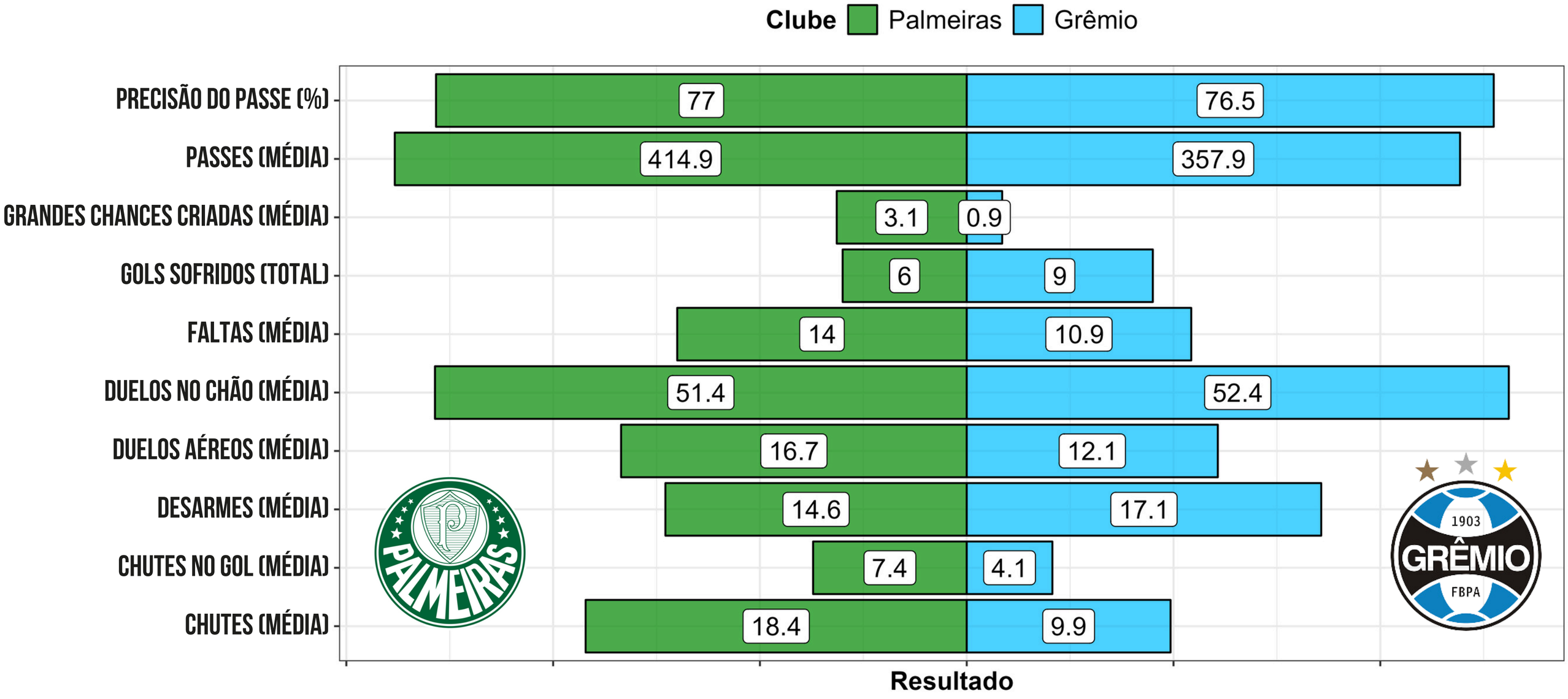
PAULO MASSINI E THM ESTATÍSTICA

THM

@THMESTATISTICA

Palmeiras x Grêmio - Análise de Desempenho

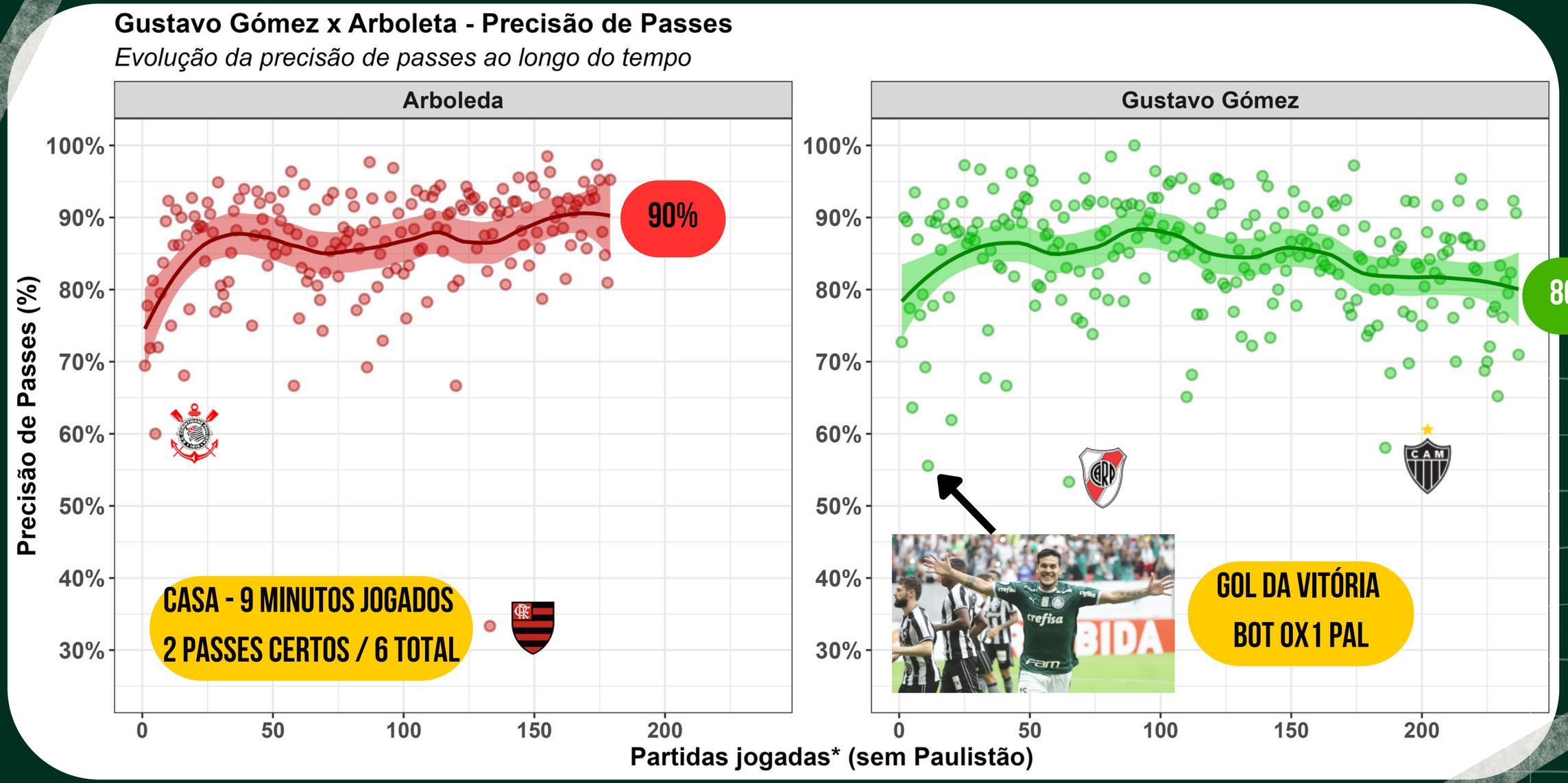
Comparação de Desempenho nos 7 últimos jogos



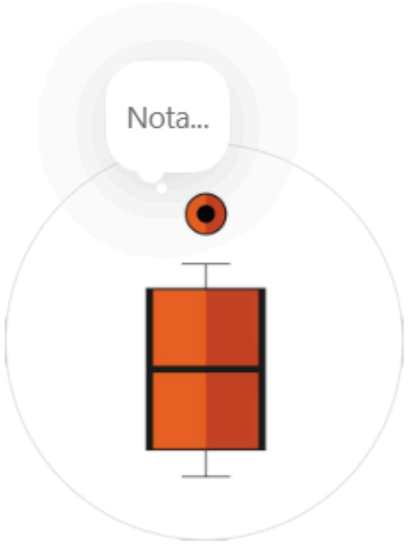
PAULO MASSINI E THM ESTATÍSTICA

ARBOLEDA X GÓMEZ

PRECISÃO DE PASSES



DADOS DE MÍDIAS SOCIAIS - INSTAGRAM E YOUTUBE



thmestatistica

Editar perfil

Itens Arquivados

527 publicações **14,5 mil** seguidores **256** seguindo

THM Estatística | Ciência de Dados

thmestatistica

Agência de consultoria

🎯 Soluções em estatística, ciência de dados e IA para alunos, pesquisadores e empresas.

🚀 +1300 alunos | 🎥 3mi views

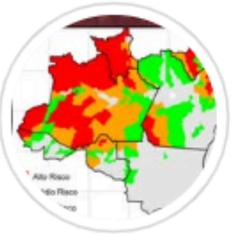
💻 ... mais

🔗 www.thmestatistica.com

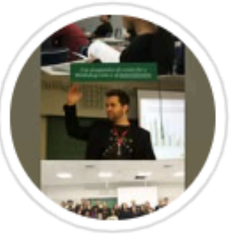
Aulas

Palestras

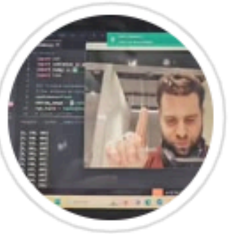
Depoimentos...



Cases THM



Workshop



Visão Comp.

Estatística 📊 ...

Obrigado!

Me mande mensagem se tiver alguma pergunta.

FIAP