**Outliers: Definição, Tratamento e Substituição em Análise de Dados**

Outliers, ou valores atípicos, são observações que se distanciam significativamente do padrão geral de uma série de dados. Frequentemente surgem em razão de erros de medição, variações naturais ou eventos excepcionais, e podem comprometer tanto a análise estatística quanto a acurácia de modelos preditivos. Em contextos aplicados, como o monitoramento de estruturas críticas, sua detecção e tratamento são essenciais.

Silva (2018) destaca que, na instrumentação de barragens, as séries temporais obtidas por sensores estão sujeitas a flutuações devido a fatores ambientais, falhas de medição e intervenções não sistemáticas. Neste cenário, outliers podem indicar desde erros pontuais (outliers aditivos) até mudanças mais estruturais, como alterações permanentes no nível da série (level shift) ou variações temporárias que se dissipam com o tempo (temporary change), conforme a tipologia proposta por Chen e Liu (1993).

A identificação desses valores anômalos pode ser realizada por métodos gráficos, como boxplots e gráficos de séries temporais, ou por técnicas estatísticas, incluindo o uso de escores Z e do intervalo interquartil (IQR) (PEÑA; TIAO; TSAY, 2001). Em aplicações mais sofisticadas, como as tratadas na monografia de Silva (2018), são empregados modelos estatísticos robustos, como ARIMA e SARIMA, que permitem detectar distorções provocadas por outliers a partir da análise dos resíduos do modelo.

O tratamento de outliers deve considerar o contexto. Em séries temporais do tipo instrumental, a remoção de dados inválidos é válida quando há evidência clara de erro. No entanto, quando os dados representam fenômenos legítimos, como variações térmicas em blocos de concreto, sua manutenção pode ser preferível. Quando a substituição é apropriada, pode-se empregar técnicas como imputação por média ou mediana, winsorização, ou ainda modelos de previsão como SARIMA, que permitem ajustar as previsões removendo os efeitos distorcidos (Silva, 2018; PEÑA et al., 2001).

É importante ressaltar que os outliers não devem ser simplesmente descartados. Sua análise pode revelar informações valiosas sobre eventos críticos, mudanças operacionais ou comportamentos emergentes nos sistemas monitorados. O estudo de Ruckstuhl et al. (2020), por exemplo, mostra como abordagens robustas de detecção de outliers são fundamentais para assegurar a integridade de análises em dados ambientais e sensoriais, como em medições de dióxido de carbono ou temperatura.

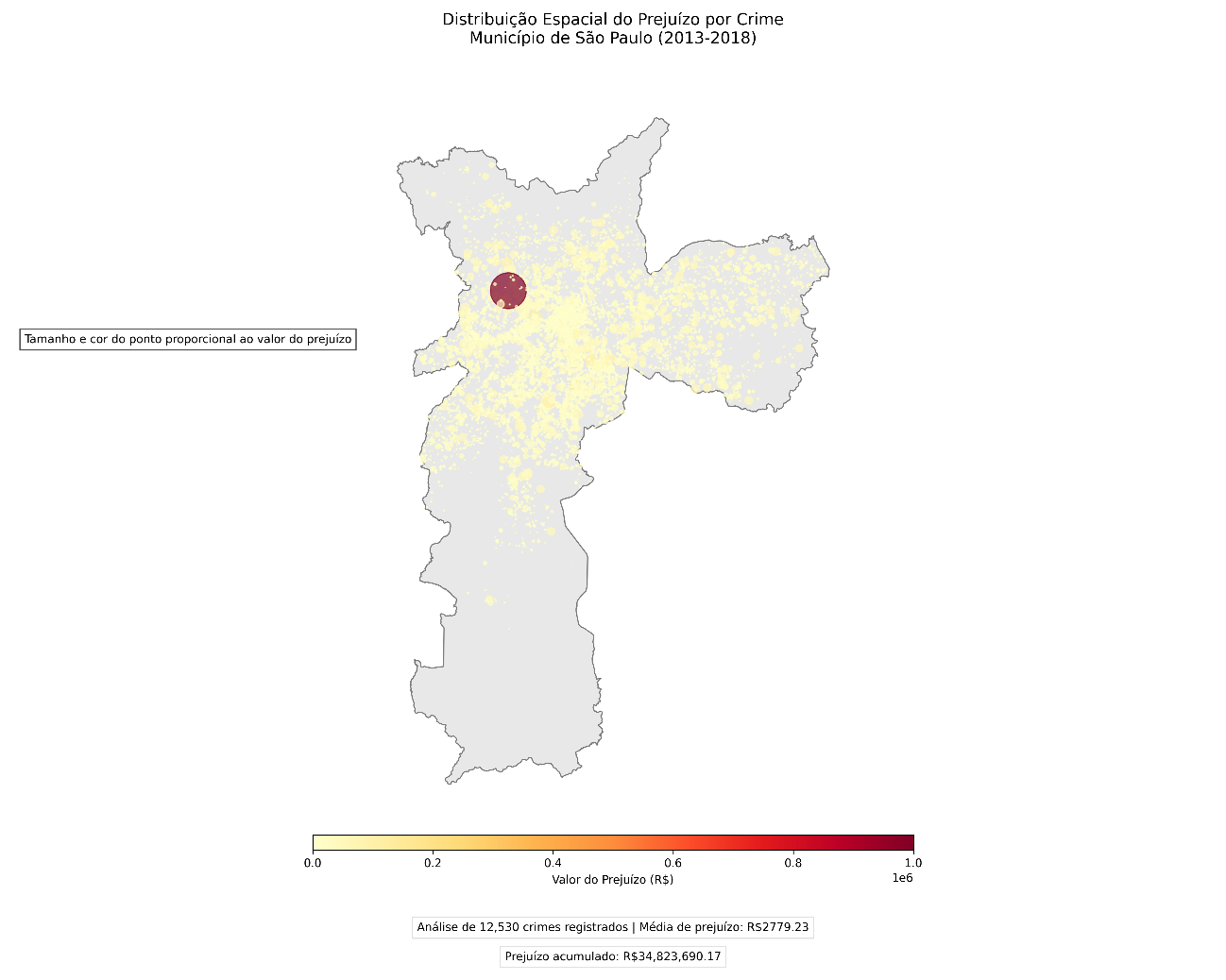
Silva (2018) demonstrou na prática que o tratamento adequado de outliers em séries temporais de instrumentos instalados em barragens contribui significativamente para a melhoria das previsões de modelos estatísticos e reduz a ocorrência de falsos alarmes em cartas de controle. Após a intervenção, observou-se um aumento na normalidade dos resíduos e uma diminuição no número de observações fora dos limites estatísticos esperados, comprovando a eficácia do método adotado.

Conclui-se, portanto, que o manejo de outliers deve ser feito com critério, levando em conta o conhecimento técnico sobre o domínio dos dados. A utilização de modelos estatísticos, aliada à experiência de especialistas, é fundamental para garantir a confiabilidade da análise e a segurança na tomada de decisões, principalmente em contextos sensíveis como o monitoramento estrutural.

Código/Mapa sem filtro de outlier:

Texto

O conteúdo gerado por IA pode estar incorreto.



O código apresentado foi retirado diretamente do seminário da disciplina de visualização de dados. Seu objetivo era mostrar a distribuição espacial do prejuízo por crime em São Paulo.

Contudo, esses dados foram retirados de um site onde o usuário insere os dados sem nenhum tipo de supervisão, visto que sua proposta seria apenas mostrar os relatos de assaltos da comunidade de São Paulo. Isso causa alguns problemas em relação a determinados valores, se observa que; uma quantidade pequena de crimes terá valores exorbitantes, uma outra parcela irá inserir valores incondizentes a realidade e por último, pode haver uma mistura dos dois cenários. Assim causando essa anomalia na visualização.

Código/Mapa Filtrado:Texto

O conteúdo gerado por IA pode estar incorreto.

Mapa

O conteúdo gerado por IA pode estar incorreto.

A solução encontrada foi calcular o percentil de 95% de valor do prejuízo e remover 5% dos maiores valores para evitar distorção na visualização e preservar a maior quantidade de dados possíveis.

Referências:

SILVA, E. A. *Detecção e Tratamento de Outliers em Séries Temporais Aplicado ao Controle Estatístico de Instrumentação de Barragens*. UFPR, 2018.

PEÑA, D.; TIAO, G. C.; TSAY, R. S. *A Course on Time Series Analysis*. Wiley, 2001.

CHEN, C.; LIU, L. Joint Estimation of Model Parameters and Outlier Effects in Time Series. *Journal of the American Statistical Association*, 1993.

RUCKSTUHL, A. et al. *Data selection methods for time series of greenhouse gases measured at background sites*. *Data in Brief*, 2020. https://doi.org/10.1016/j.dib.2020.106276

Links usados: <https://www.sciencedirect.com/science/article/pii/S2352340920311707?via%3Dihub>

<https://www-sciencedirect-com.translate.goog/science/article/abs/pii/S2352409X20302145?_x_tr_sl=en&_x_tr_tl=pt&_x_tr_hl=pt&_x_tr_pto=wa>

**Pesquisa de Silva disponível no .rar**