Influência de Outliers na Identificação de Processos

Fidel Ernesto Díaz Andino * **Jose Angel Medel Tirador** ** **Claudio Garcia** ***

* *Departamento de Engenharia de Telecomunicações e Controle. Escola Politécnica da Universidade de São Paulo, USP, SP.(e-mail: fediaza@usp.br)*
** *Departamento de Engenharia de Telecomunicações e Controle. Escola Politécnica da Universidade de São Paulo, USP, SP.(e-mail: joseangel93@usp.br)*
*** *Departamento de Engenharia de Telecomunicações e Controle. Escola Politécnica da Universidade de São Paulo, USP, SP.(e-mail: clgarcia@lac.usp.br)*

Abstract: One of the main problems of data acquisition in some industrial processes is the process noise that sometimes generates outliers in the collected data. If such data were to be used in process identification, the analysis and information gathering could be impaired. The purpose of this work is to study the difference between the identification of raw data and "outlier-clean" data, by modeling and comparing the performance of these models. To achieve it, there were created over 200 simulated plants. At the end, as a validation method, the experiments were also conducted in a real plant.

Resumo: Um dos principais problemas de aquisição de dados em alguns processos da indústria é o ambiente ruidoso que às vezes gera outliers nos dados coletados. Se tais dados forem usados para identificar o processo, a análise e coleta de informações poderiam ser prejudicadas, fazendo com que a modelagem seja ineficiente. O propósito deste trabalho é realizar uma comparação da modelagem dos dados crus, isto é, sem processar (com outliers), e os dados "limpos", mediante alguns índices de desempenho. Para isso, foram simuladas 200 plantas. No final, como forma de validação, os experimentos foram feitos com dados coletados de uma planta real.

Keywords: Outliers, System identification.
Palavras-chaves: Outliers, Identificação de sistemas.

1. INTRODUÇÃO

A sociedade atual é muito dependente de operações automáticas em uma grande quantidade de processos, alguns tão simples como a temperatura ou velocidade de rotação de um motor, outros tão complexos como o voo de um avião. Pequenos processos que ao falhar poderiam causar a morte de pessoas ou grandes que poderiam causar eventos catastróficos. Algo tão simples como a queda no desempenho poderia ter efeitos indesejados até na economia de um país, pelo que seu monitoramento e controle se torna de grande importância, tanto para garantir a qualidade, bem como para manter o desempenho.
Com o passar do tempo, os processos se tornam mais complexos devido às altas demandas da uma sociedade consumidora, pelo que a complexidade de mecanismos para garantir o controle adequado tem crescido muito. Assim novos desafios surgem a cada dia, pois nem sempre se tem conhecimento suficiente sob o processo, por exemplo, para descrevê-lo usando modelos matemáticos. Como resultado, o uso de dados coletados por sensores para gerar modelos matemáticos através da Identificação de Sistemas, tornou-se uma solução atraente em contraste com as técnicas clássicas existentes de modelagem fenomenológica.
Do ponto de vista das aplicações práticas, esses dados podem estar afetados por ruídos e outliers, que podem ser um grande fator de deterioração da qualidade dos modelos gerados por identificação de sistemas.
Na literatura, diversos autores propõem diferentes definições para tais valores discrepantes, tais como anomalias, observações discordantes, surpresas, exceções, aberrações, e outras, mas nenhuma das anteriores é um conceito formal. Uma das definições mais usadas para outliers, foi a proposta por Barnett e Lewis (1994):

"Uma observação (ou subconjunto de observações) que parece ser inconsistente com o restante desse conjunto de dados"

Pelo que, pode-se resumir como: observações que não seguem o comportamento esperado dos dados. Estas estão relacionadas com erros que podem ser mecânicos, ambientais, digitais, de software, e outros que incluem os erros causados por humanos.
No contexto de controle de processos, outliers podem estar presentes tanto em observações na saída quanto na entrada do processo, devido a falhas de sensores, erros de transmis-

são de dados, ou outros fatores. Esses dados contaminados podem levar a uma representação equivocada da relação entre as variáveis de entrada e saída, obtidas por qualquer método de identificação de sistemas.
Neste contexto, surgem duas alternativas para solucionar o problema, a primeira e mais comum, refere-se a um método de dois passos: o primeiro passo é a detecção dos outliers, e o segundo, é a remoção ou substituíção dos mesmos por valores mais razoáveis; logo com os dados resultantes identificar os modelos. A segunda alternativa refere-se ao uso de alguns algoritmos para identificar sem a prévia detecção/remoção dos outliers.
O propósito do presente trabalho é realizar um estudo sobre a influência dos outliers na qualidade de modelos obtidos através da segunda alternativa. Para isso, o trabalho está dividido da seguinte forma: na Seção 2, tenta-se formalizar de forma resumida o problema dos outliers em dados coletados, na Seção 3 citam-se modelos e métodos usados na identificação, a Seção 4, faz referência aos experimentos feitos e por ultimo, na Seção 5 se tira uma conclusão sobre os resultados obtidos.

2. FORMALIZAÇÃO DO PROBLEMA

A identificação de sistemas está relacionada as técnicas usadas para estudar um processo por meio de dados observados/medidos, com o objetivo de obter uma relação matemática (modelo) entre as entradas e saídas do processo. A expectativa é que o modelo encontrado consiga descrever as dinâmicas do processo o mais próximo possível (Ljung, 1999).

2.1 Observações

O conjunto de observações coletadas é conhecido como serie temporal, por serem coletadas sequencialmente no tempo. A classificação de acordo com o número de variáveis (temperatura, nível, pressão,...) observadas pode se dividir em univariada ou multivariada. Formalizando, pode-se definir como (Karimi-Bidhendi et al., 2018):

Definição *1.* Uma série temporal univariada é definida como $(y_t)_{t=1}^{N}$, na qual y_t é a observação, t é o instante de coleta e N o número de observações que compõem a série.

Definição *2.* Uma série temporal multivariada é definida como $(Y_t)_{t=1}^{N}$, na qual $Y_t = [y_{t1}, y_{t2}, ..., y_{tM}]$ é um arranjo de séries univariadas, com t sendo o instante de coleta e N o número de variáveis que compõem a série. Neste caso, as variáveis podem depender não só de seus valores anteriores mas também das outras variáveis da série (no caso de processos, entradas e saídas).

Por ser o sistema mais simples, composto por dois sinais (uma entrada e uma saída), a partir deste momento estará se fazendo referência a séries temporais multivariadas.

2.2 Outliers

Neste âmbito foram estudados dois tipos de outliers que são caraterizados como se segue (Chandola; Kumar, 2009):

Definição *3. Ponto discordante* (conhecido como tipo 1 na literatura): quando um ponto do conjunto de dados se desvia do padrão normal do conjunto (Figura 1).

Exemplo deste tipo de outlier na indústria, é aquele gerado por erros na hora de transmitir a informação dos diversos medidores até o sistema aonde são armazenados os dados. Podem ser causados por interferência ambiental (ruído, choque, vibração,...), obstáculos de comunicação, falha de sensores, entre outros fatores (Zhang et al., 2010).

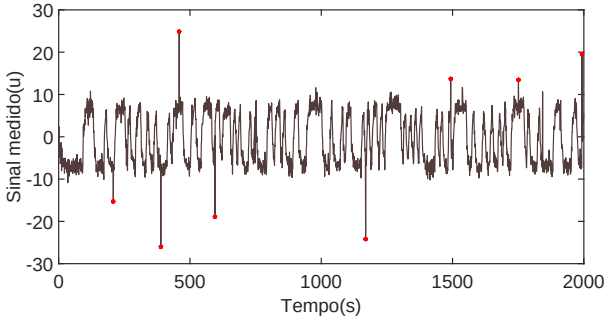

Figura 1. Exemplo de observações coletadas na saída de um processo contaminado com outliers do tipo 1, marcados em vermelho.

Definição *4. Conjunto discordante* (conhecido como tipo 2 na literatura): quando um conjunto de dados está se comportando de modo anormal em relação ao resto do conjunto (Figura 2).

Pode ocorrer que a discordância não seja uma anomalia por si só, mas por pertencer a outro conjunto, ela é identificada como tal. Esse tipo de outlier pode-se encontrar em um processo por causa da interferência de um sistema desconhecido ou não esperado.

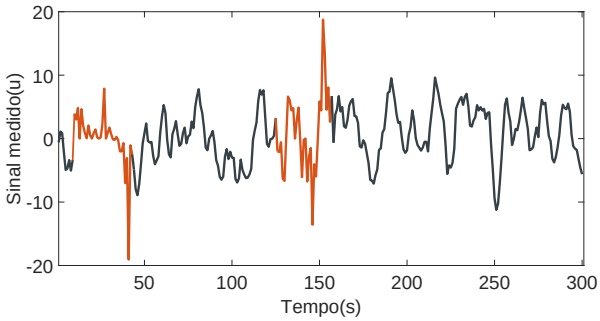

Figura 2. Exemplo de observações coletadas na saída de um processo contaminado com outliers do tipo 2, conjunto marcado em laranja.

Uma maneira comum de detectar outliers é por simples inspeção visual dos gráficos de dados com base na experiência dos técnicos. Esse método é impreciso e é ineficaz em muitos casos, por causa da complexidade dos sistemas e das grandes quantidades de dados que são geradas a cada segundo. A Figura 2 é um exemplo claro disso.

3. MODELOS E MÉTODOS DE IDENTIFICAÇÃO

Comumente a identificação de modelos empíricos se faz sem conhecimento do processo, significando que se desconhece tanto a estrutura quanto qualquer aproximação dos parâmetros. Isto se conhece como identificação caixa preta (black-box). O caso em que se conhece alguma característica do processo, ou se quer impor uma estrutura ou algumas restrições sobre esta (ordem, quantidade de parâmetros,...), é conhecido como identificação caixa cinza.

Por motivos práticos, neste trabalho será usado o primeiro método nos experimentos. No contexto de identificação serão utilizados sistemas lineares invariantes no tempo (SLIT) e os supostos que isso assume [1].

3.1 Modelos e estimadores

Para identificar um modelo é preciso ter um conjunto de estruturas de modelos, um critério ou função perda (como a Equação (3), definida um pouco mais adiante), para selecionar os melhores modelos, e uma regra para avaliar os modelos candidatos (ver Subseção 3.4), com base nos dados (Ljung, 1999). Exemplos de estruturas de modelos são equações em espaços de estados com matrizes do sistema desconhecidas, funções de transferências com polos, zeros e ganho ajustáveis ou funções parametrizáveis. A Equação (1) representa uma estrutura simples de modelo, onde b_1 e b_0 são parâmetros ajustáveis.

$$y_i = b_1 x_i + b_0 \tag{1}$$

Alguns dos modelos mais conhecidos na literatura de controle de processos são obtidos através do representado na Equação (2), chamado de modelo geral linear.

$$A(q)y(k) = \frac{B(q)}{F(q)}u(k-n) + \frac{C(q)}{D(q)}e(q) \tag{2}$$

onde A, B, C, D e F são polinômios, $y(k)$, $u(k)$, $e(k)$ e n as saídas, entradas, perturbação e o atraso do sistema, e q é o operador de deslocamento.

A seguir se apresentam as estruturas mais usadas e que fazem parte do presente trabalho:

(1) Modelo auto regressivo com entrada exógena (ARX): Derivado do modelo linear geral assumindo $C(q) = D(q) = F(q) = 1$;
(2) Média móvel auto regressiva com entrada exógena ($ARMAX$): Derivado do modelo linear geral assumindo $D(q) = F(q) = 1$;
(3) Erro de saída (OE): Esta estrutura não inclui um modelo de perturbação, é derivada do modelo linear geral assumindo $A(q) = C(q) = D(q) = 1$;
(4) Box-Jenkins (BJ): Derivada da estrutura geral, assumindo $A(q) = 1$.

3.2 Métodos

Após escolhidas as estruturas dos modelos, são empregados métodos baseados em regressão para estimar os parâmetros. Um dos métodos mais conhecidos e usados é o MQ (Mínimos Quadrados). Como demonstração, a seguir se apresenta a estimação dos parâmetros b_0 e b_1 de (1), usando o estimador MQ (Greene, 2013). Primeiramente:

(1) Define-se a estrutura e parâmetros a estimar, no caso (1), a variável estimada se denota como $\hat{y}$, logo,
(2) Se minimiza a soma dos quadrados do erro (SQE), onde:

$$SQE = \sum_{i=1}^{n}(y_i - \hat{y})^2 = \sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2 \tag{3}$$

Calculando e minimizando se obtém o resultado mostrado em (4).

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, b_0 = \frac{\sum_{i=1}^{n} y_i}{n} - b_1 \frac{\sum_{i=1}^{n} x_i}{n} \tag{4}$$

Só falta avaliar se o modelo estimado se ajusta de forma adequada aos dados reais e escolher, de acordo com a avaliação, qual é o melhor modelo. Neste trabalho se usaram vários critérios que são expostos na Subseção 3.4.

3.3 Sensibilidade à presencia de outliers

Uma medida da sensibilidade de um estimador a outliers é o *breakdown point* (BP), que essencialmente quantifica o quão bem uma estimativa pode suportar dados "ruins" (Donoho; Huber, 1983). Para amostras finitas, a definição é a menor fração de contaminação que pode fazer com que o estimador produza estimativas arbitrariamente maiores do que aquelas obtidas a partir dos dados não contaminados (Donoho; Huber, 1983). Basicamente, isso significa que quanto menor o BP de um estimador, menor será sua robustez frente a outliers.

Estimadores clássicos como média e variância são altamente sensíveis à presença de outliers, tendo estes um BP de 0%. Apesar de sua simplicidade matemática e baixo custo computacional, o estimador MQ também é pouco robusto, pois um único outlier pode ter um efeito muito grande na estimativa, distorcendo completamente a função estimada. Este estimador também tem um BP de 0%.

Devido a estas razões foram surgindo métodos mais robustos como o LMS (Least Median of Squares) (Rousseeuw, 1984), o LTS (Least Trimmed Squares) (Mount et al., 2014), e outros, que são inerentemente menos sensíveis a outliers, alguns com um BP que se aproxima de 50%, que por definição é o maior BP que pode alcançar um estimador (Donoho; Huber, 1983). Por isso se estará fazendo uso de tais estimadores robustos neste trabalho, em específico o estimador LTS.

3.4 Índices para avaliar os Modelos

No contexto deste trabalho, mais que avaliar os modelos, os índices aqui apresentados são utilizados para realizar uma comparação entre os modelos obtidos a partir da contaminação ou não de cada tipo de outlier. Os índices usados são apresentados a seguir:

(1) *FIT (Fitness).* O índice é um indicador de quão boa é a resposta do modelo estimado em comparação com os dados medidos. O fit aqui utilizado é definido como a porcentagem do índice NRMSE (Normalized Root Mean Square Error) que é dado por $NRMSE = (1 - \frac{1}{N}\sum_{k=1}^{N}\frac{||y_k - \hat{y}_k||_2}{||y_k - E(y_k)||_2})$, onde y_k e $\hat{y}_k$ são a saída do sistema e do modelo respectivamente, e $E(y_k)$ é o valor esperado (Prívara et al., 2013), logo o índice apresentado é $fit = 100 * NRMSE(\%)$. Um índice mais próximo de 100% indica um modelo melhor.
(2) *TIC (Theil's Inequality Coefficients)*: Os coeficientes de Theil foram projetados para medir a precisão de um conjunto de dados gerados por um modelo. Theil propôs dois índices para a tarefa. O índice U_1 aqui utilizado, oferece uma pontuação que está no intervalo de [0;1]. Um valor de 0 indica uma predição perfeita enquanto um valor de 1 corresponde a uma perfeita

[1] Superposição e Invariância no Tempo

desigualdade entre os dados preditos e os medidos (Lehthold, 1975). Uma definição mais completa dos coeficientes pode ser encontrada em Theil (1966).

(3) AIC (Akaike Information Criterion): O critério fornece uma medida da qualidade do modelo, simulando o modelo com conjuntos de dados diferentes. Este critério pode ser usado para comparar vários modelos. De acordo com o critério de Akaike, o modelo mais preciso tem o menor AIC (Akaike, 1974).

4. EXPERIMENTAÇÃO

Como parte da experimentação, foram simulados vários sistemas (processos) criados aleatoriamente, para os quais se tinha disponível um modelo. Esses sistemas foram excitados e as saídas coletadas, simulando um experimento de identificação real. Os dados coletados passaram por um processo de "contaminação" de outliers, para o primeiro experimento do tipo 1, para o segundo do tipo 2 e por último com os dois tipos. O processo de "contaminação" é descrito nos algoritmos (1) e (2):

Algoritmo 1: Infecta um sinal com pontos discordantes.

```
1: function InfectaComTipo1(sinal)
2:     ymax ← 2Max( abs(sinal) )
3:     ymin ← -2Max( abs(sinal) )
4:     ylength ← Length(sinal)
5:     locout[ ] ← IntegerRand(0.05 · ylength)
6:     outiers[ ] ← IntegerRand([ymin ymax], locout)
7:     olocation ← RandPerm(outiers, ylength)
8:     sinal(olocation) ← sinal(olocation) + outiers
9: return sinal
```

Algoritmo 2: Infecta um sinal com conjuntos discordantes.

```
1: function InfectaComTipo2(sinal)
2:     ymax ← 2Max( abs(sinal) )
3:     ymin ← -2Max( abs(sinal) )
4:     locout[ ] ← IntegerRand( 0.005 · Length(sinal) )
5:     rankmax ← 10 · IntegerRand([5 10])
6:     rank[ ] ← IntegerRand([20 rankmax])
7:     outliers[ ] ← IntegerRand([ymin ymax], rank)
8:     for i=1:Length(locout) do
9:         aux[ ] ← locout(i) : (locout(i) + rank(i))
10:        sinal(aux) ← sinal(aux). ∗ outliers
11: return ytipo2
```

Os algoritmos apresentados colocam outliers das definições (3) e (4) respectivamente, em posições aleatórias no sinal passado como parâmetro, e retornam o sinal modificado. Depois de gerar os dados, se procedeu com a identificação e validação dos modelos. Foram simulados um total de 200 modelos e se estimaram os índices estatísticos de interesse. Para maior entendimento, a seguir se apresenta o resultado de uma simulação.

4.1 Resultado de uma simulação

O modelo obtido (sistema) para a simulação escolhida se apresenta em (5).

$$y[k] = \frac{1.658q^{-1} - 0.4349q^{-2}}{1 - 0.7104q^{-1} + 0.1111q^{-2}}u[k] \qquad (5)$$

Para excitar o sistema foi criado um sinal PRBS [2], com 2000 amostras de comprimento e níveis que estão no intervalo de $[-1; 1]$, e um período de amostragem $T_s = 1$. A Figura 3 mostra o sistema simulado com outliers do tipo 1, o sinal PRBS é mostrado no sub-gráfico 3.

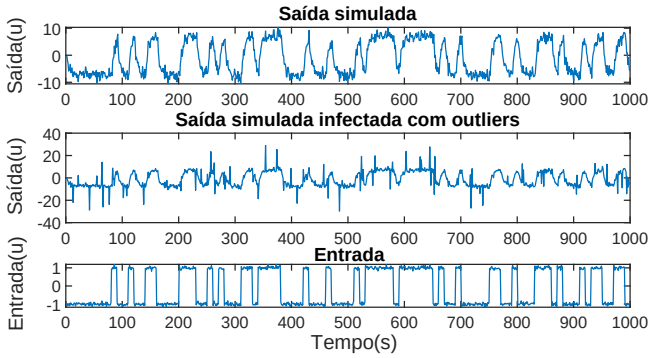

Figura 3. Simulação do sistema (Equação 5), sob a influência de outliers do tipo 1 com $T_s = 1$.

A identificação do sistema final, isso é, contendo outliers, gerou os seguintes resultados, $Fit = 73.53\%$, $TIC = 0.6725$, $AIC = 1016$. Note que os dados mostrados são a mediana (medida robusta) dos resultados obtidos, da mesma forma são apresentados o resto dos resultados. Comummente os dados são separados em dois conjuntos, com a ideia de usar um para identificar modelos que melhor descrevam o processo e outro para validar os modelos encontrados. Neste caso, foram gerados dados novos para validar, pelo que tal separação não foi necessária.

A Figura 4 mostra uma comparação dos modelos obtidos com o sistema original (antes da inserção dos outliers). Note como o método de estimação foi capaz de obter bons modelos, apesar de não se ter dados limpos (livre de outliers), isso é devido ao alto BP do estimador usado.

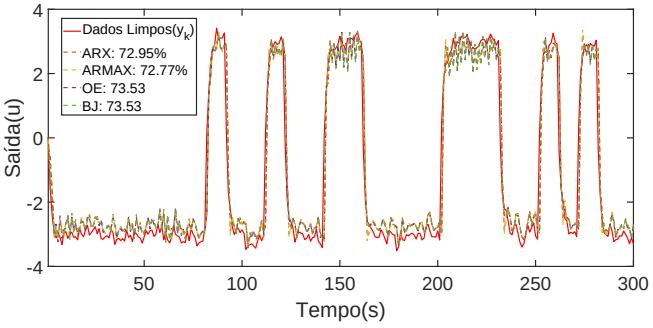

Figura 4. Comparação da planta (Equação 5) com os modelos estimados, sob a influência de outliers do tipo 1.

Um dos modelos obtidos é apresentado em (6).

$$y[k] = \frac{2.199q^{-1} + 1.843q^{-2}}{1 + 0.7102q^{-1} - 0.2528q^{-2}}u[k] \qquad (6)$$

Note que o modelo tem a mesma estrutura mas os parâmetros estão longe de ser os reais, no entanto as análises feitas sobre o modelo falam por si só, e demonstram que é bastante aceitável para ser um modelo estimado sobre dados perante a influência de outliers. Resultados semelhantes são os apresentados na Tabela 1, que mostra os índices obtidos para o resto dos experimentos deste tipo.

[2] Sequência binária pseudoaleatória.

Tabela 1. Índices estatísticos obtidos para todos os modelos estimados. Contaminação do tipo 1.

Estrutura	Fit (%)	TIC	AIC
ARX	69.33	0.6791	1106
ARMAX	69.84	0.6877	1076
OE	63.43	0.7099	1298
BJ	72.16	0.6827	1057

Da mesma forma, foram realizados experimentos com dados contaminados com outliers do tipo 2. A Figura 5 mostra a saída do sistema após contaminação, vale enfatizar que o sinal usado é o mesmo obtido a partir da excitação do sistema no experimento anterior. Nenhum gráfico mostra o comprimento total do sinal, por uma questão de espaço e visibilidade, pois, como especificado no começo da Seção 4, o sinal de excitação é de 2000 amostras e o T_s é de 1 segundo, pelo que o sinal de saída tem o mesmo comprimento.

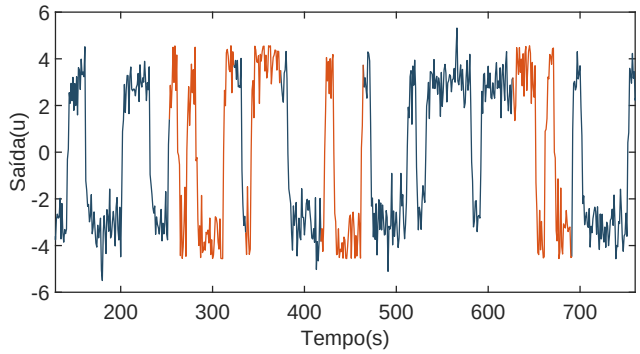

Figura 5. Simulação do sistema (Equação 5), sob a influência de outliers do tipo 2 com $T_s = 1$.

Note que a dinâmica do sistema aparentemente é a mesma, o que demonstra que não é suficiente fazer uma detecção/limpeza dos dados por simples inspeção visual. Novamente se repete o processo de identificação e validação do modelo, para o qual são obtidos os seguintes resultados: $Fit = 29.64\%$, $TIC = 0.9549$, $AIC = 5549$.

A Figura 6 mostra a comparação dos modelos obtidos com os dados limpos. Desta vez os modelos tiveram um ajuste pobre, e os índices calculados são muito inferiores aos calculados no experimento anterior.

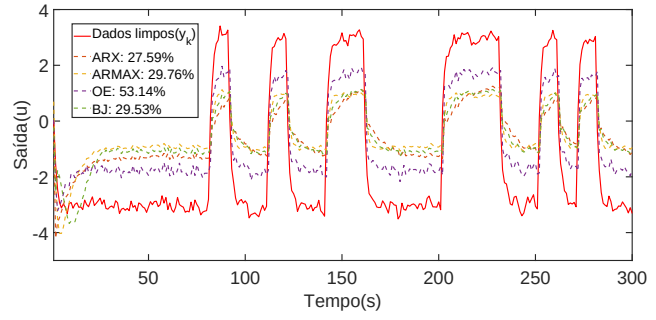

Figura 6. Comparação da planta (Equação 5) com os modelos estimados, sob a influência de outliers do tipo 2.

As simulações restantes seguiram o mesmo comportamento, gerando resultados ruins como evidenciado na Tabela 2, o que sugere que este tipo de outliers influencia na estimação dos modelos.

Tabela 2. Índices estatísticos obtidos para os modelos estimados. Contaminação do tipo 2 (T_2) e contaminação dos tipos 1 e 2 ($T_{1,2}$).

	Fit(%)		**TIC**		**AIC**	
Estrutura	T_2	$T_{1,2}$	T_2	$T_{1,2}$	T_2	$T_{1,2}$
ARX	46.23	44.83	0.8120	0.8172	1891	1900
ARMAX	48.38	45.87	0.8049	0.8187	1856	1890
OE	45.03	43.93	0.7942	0.8125	1886	1966
BJ	47.36	45.96	0.8020	0.8116	1867	1892

Note ainda a pequena diferença entre os índices calculados para os modelos obtidos nos experimentos 2 e 3. Isso demonstra novamente que outliers do tipo 1 têm pouca ou nenhuma influência sobre a estimação.

4.2 Planta real

Na Figura 7 se mostra o P&ID simplificado de uma planta de neutralização de pH. A planta é composta por três tanques primários aonde se armazenam as soluções, ácido ($TA_{1,2}$) e base (TB), e um tanque principal, o tanque reator (TR), aonde acontece o processo de mistura de ambas as soluções, com ajuda de um agitador mecânico. A vazão de todas as entradas no TR é medida por medidores magnéticos (M), e o nível é controlado por uma válvula, que regula a vazão de saída. O TR conta também com um pH-metro para monitorar o pH da reação, e um conjunto sensor/resistência para controlar a temperatura do processo. A vazão de saída do TB é manipulada por uma bomba. Os reagentes empregados são ácido clorídrico (HCL) e hidróxido de sódio (NaOH).

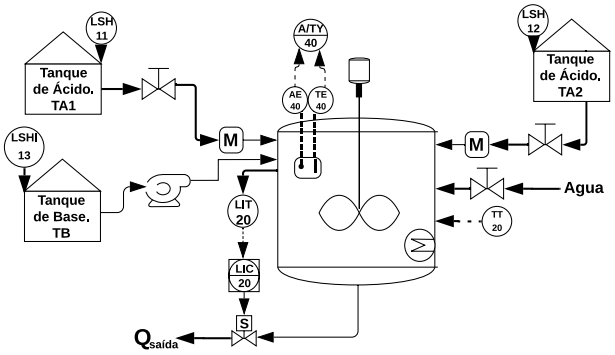

Figura 7. Planta de neutralização de pH [adaptado de (Garcia; Juliani, 2011)].

Embora o processo tenha várias variáveis que poderiam ser parte do experimento, decidiu-se utilizar o nível por sua simplicidade. O procedimento para a identificação foi o mesmo anteriormente comentado. Desta vez, a contaminação por outliers aconteceu da seguinte forma: para os outliers do tipo 1, foi manuseado o sensor da variável escolhida (LIT), falseando assim algumas medições. Já para o tipo 2, foi mexido no TR criando ondas na solução por determinado tempo. Após feita a análise para os dados contaminados com outliers do tipo 1, chegou-se ao mesmo resultado alcançado nas simulações realizadas, como se mostra na Tabela 3, coluna T_1.

A Figura 8 mostra o resultado do experimento 2, já com os modelos identificados. Note que no sinal coletado são marcadas com setas e elipses em vermelho, algumas

das contaminações efetuadas. O objetivo é poder ver a influência causada pelos outliers de forma gráfica.

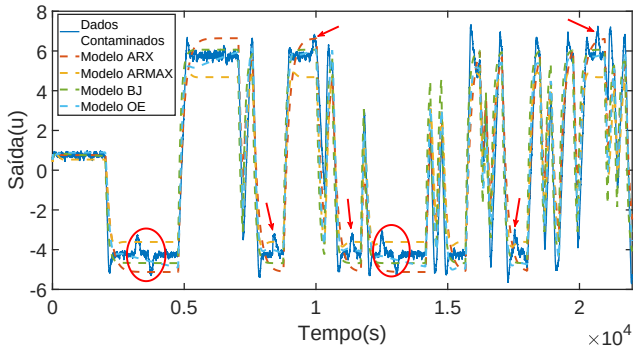

Figura 8. Modelos obtidos para a planta real sob a influência de outliers do tipo 2.

A Tabela 3 mostra também os índices estatísticos dos modelos estimados para a contaminação tipo 2 (coluna T_2). Não são inclusos os resultados obtidos pela mistura de ambos, por causa da grande semelhança com os resultados do experimento 2 (coluna T_2).

Tabela 3. Índices estatísticos obtidos para os modelos estimados. Contaminação dos tipos 1 e 2.

	Fit(%)		**TIC**		**AIC**	
Estrutura	T_1	T_2	T_1	T_2	T_1	T_2
ARX	71.11	66.21	0.1447	0.1576	-85107	-84512
ARMAX	73.51	68.68	0.1344	0.1737	-85500	-85809
OE	73.76	66.13	0.1335	0.1590	65896	67272
BJ	73.47	65.82	0.1351	0.1687	-85388	-3689

5. CONCLUSÃO

Existem duas alternativas para resolver o problema da identificação de sistemas em presencia de outliers, a primeira e mais comum, é aquela baseada em algoritmos de detecção/correção dos outliers para uma posterior identificação. O problema com esta alternativa, é que depende do usuário para definir alguns parâmetros que irão influenciar no resultado. Por exemplo em técnicas baseadas em "janelas deslizantes" deve-se definir o tamanho da janela, ou tentar encontrá-lo por outros meios. Se o tamanho definido for errado, os resultados podem não ser os esperados.
Neste trabalho se aborda a segunda alternativa existente, isto é, o uso de algoritmos robustos para a estimação. Este campo de estudo, apesar de remontar a vários anos atrás, ainda não é muito explorado. No estudo realizado, se infectaram vários conjuntos de dados com outliers, combinando o efeito destes para estudar como se comportam os algoritmos deste tipo. Como parte da pesquisa se fez uma avaliação dos modelos obtidos usando esta técnica, por meio de uma comparação com os modelos reais. Foi evidenciado que usando esta alternativa, pode-se obter resultados que vão depender meramente da porcentagem de corrupção dos dados.
Os resultados alcançados determinaram que, se o estimador usado for robusto, outliers do tipo 1 não influenciam a estimação do modelo. Esta afirmação é verdadeira até certo ponto, pois quando a quantidade de outliers se aproxima de 50% do sinal, nem estimadores robustos conseguem fazer uma boa estimação. Para outliers do tipo 2, se evidenciou uma diferença às vezes notável com respeito aos sistemas reais. Isso é devido a que estes conjuntos discordantes de fato mudam a dinâmica do sinal, aderindo o novo comportamento a quase todos os modelos estimados. Quando misturados ambos os tipos (1 e 2), os resultados obtidos são semelhantes aos da contaminação do tipo 2.
Então se pode concluir que em caso de dados contaminados por outliers do tipo 1, é mais factível identificar usando estimadores robustos, já para o tipo 2 (ou quando misturados), deve-se decidir entre esta alternativa ou a primeira, ou inclusive as duas.

REFERÊNCIAS

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.

Chandola, V. e Kumar, V. (2009). Outlier detection: A survey. *ACM Computing Surveys*, 41(3).

Donoho, D.L e Huber, P. (1983). The notion of breakdown point. *In: Bickel PJ, Doksum K, Hodges JL Jr (eds) A Festschrift for Erich L. Lehmann. Wadsworth, Belmont, CA*, pp 157–184.

Garcia, C. e Juliani, R. (2011). Modelling and Simulation of pH Neutralization Plant Including the Process Instrumentation. *Applications of MATLAB in Science and Engineering*, 485 – 510. doi:10.5772/24718.

Greene, N. (2013). Generalized least-squares regressions I: Efficient derivations. *Proceedings of the 1st International Conference on Computational Science and Engineering (CSE '13)*, Valencia, Spain, August 6-8, 2013.

Karimi-Bidhendi, S., Munshi, F., e Munshi, A. (2018). Scalable classification of univariate and multivariate time series. In *2018 IEEE International Conference on Big Data (Big Data)*, 1598–1605. doi:10.1109/BigData.2018.8621889.

Lehthold, R.M. (1975). On the use of theil's inequality coefficients. *American Journal of Agricultural Economics*, 57, 344–346.

Ljung, L. (1999). System Identification: Theory for the User, 2nd Edition. *Prentice Hall PTR*.

Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., e Wu, A.Y. (2014). On the least trimmed squares estimator. *Algorithmica*, 69(1), 148–183. doi:10.1007/s00453-012-9721-8.

Prívara, S., Cigler, J., Váňa, Z., Oldewurtel, F. e Žáčeková, E. (2013). Use of partial least squares within the control relevant identification for buildings. *Control Engineering Practice*, 21(1), 113–121. doi:10.1016/j.conengprac.2012.09.017.

Rousseeuw, P. (1984). Least median of squares regression. *Journal of The American Statistical Association - J AMER STATIST ASSN*, 79, 871–880. doi:10.1080/01621459.1984.10477105.

Theil, H. (1966). Applied economic forecasting. *North-Holland Publising Company*.

Zhang, Y., Meratnia, N., e Havinga, P. (2010). Outlier detection techniques for wireless sensor networks: A survey. *IEEE Communications Surveys Tutorials*, 12(2), 159–170. doi:10.1109/SURV.2010.021510.00088.