Input Sensitivity on the Performance of Configurable Systems: An Empirical Study
Luc Lesoil, Mathieu Acher, Arnaud Blouin, Jean-Marc Jézéquel

To cite this version:

HAL Id: hal-03476464
https://inria.hal.science/hal-03476464v2
Submitted on 16 Feb 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Input Sensitivity on the Performance of Configurable Systems: An Empirical Study

Luc Lesoil\textsuperscript{a}, Mathieu Acher\textsuperscript{a}, Arnaud Blouin\textsuperscript{a}, Jean-Marc Jézéquel\textsuperscript{a}

\textsuperscript{a}Univ Rennes, Inria, INSA Rennes, CNRS, IRISA, France

Abstract

Widely used software systems such as video encoders are by necessity highly configurable, with hundreds or even thousands of options to choose from. Their users often have a hard time finding suitable values for these options (i.e., finding a proper configuration of the software system) to meet their goals for the tasks at hand, e.g., compress a video down to a certain size. One dimension of the problem is of course that performance depends on the input data: e.g., a video as input to an encoder like \textit{x264} or a file fed to a tool like \textit{xz}. To achieve good performance, users should therefore take into account both dimensions of (1) software variability and (2) input data. This paper details a large study over 8 configurable systems that quantifies the existing interactions between input data and configurations of software systems. The results exhibit that (1) inputs fed to software systems can interact with their configuration options in non-monotonous ways, significantly impacting their performance properties (2) input sensitivity can challenge our knowledge of software variability and question the relevance of performance predictive models for a field deployment. Given the results of our study, we call researchers to address the problem of input sensitivity when tuning, predicting, understanding, and benchmarking configurable systems.

Keywords: Configurable Systems, Input Sensitivity, Performance Prediction

1. Introduction

Widely used software systems are by necessity highly configurable, with hundreds or even thousands of options to choose from. According to Svahnberg \textit{et al.} \cite{svahnberg90}, software variability is the "ability of a software system or artifact to be efficiently extended, changed, customized or configured for use in a particular context". For example, a tool like \textit{x264} offers 118 run-time options such as --ref, --no-mbtree or --no-cabac for encoding a video. The same applies to \textit{Linux} kernels or compiler such as \textit{gcc}: they all provide configuration options through compilation options, feature toggles or command-line parameters. Software engineers often have a hard time finding suitable values for those options (i.e., finding a proper configuration of the software system) to meet their goals for the tasks at hand, e.g., compile a program into a high-performance binary or compress a video down to a certain size while keeping its perceived quality.

Since the number of possible configurations grows exponentially with the number of options, even experts may end up recommending sub-optimal configurations for such complex software \cite{40}. However, there exists cases where inputs (e.g., files fed to an archiver like \textit{xz} or SAT formulae provided as input to a solver like \textit{lingeling}) can also impact software variability \cite{105,54}. The \textit{x264} encoder typifies this problem, as illustrated in Figure 1. For example, Kate, an engineer working for a VOD company, wants \textit{x264} to compress input videos to the smallest possible size. She executes \textit{x264} with two configurations \textbf{C} (with options --no-mbtree --ref 1) and \textbf{C’} (with options --no-cabac --ref 16) on the input video \textit{I}_1 and states that \textbf{C} is more appropriate than \textbf{C’} in this case. But when trying it on a second input video \textit{I}_2, she draws opposite conclusions: for \textit{I}_2, \textbf{C’} leads to a smaller output size than \textbf{C}. Now, Kate wonders what configuration to choose for other inputs, \textbf{C} or \textbf{C’}? More generally, do configuration options have the same effect on the output size despite a different input? Do options interact in the same way no matter the inputs? These are crucial practical issues: the diversity of existing inputs can alter her knowledge of \textit{x264}’s variability. If it does, Kate would have to configure \textit{x264} as many times as there are inputs, making her work really tedious and difficult to automate.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{fig1.png}
\caption{This paper explores and quantifies how inputs fed to software systems impact performance of configurations.}
\end{figure}
for a field deployment.

Numerous research works have proposed to model performance of software configurations, with several use-cases in mind for developers and users of software systems: the maintenance and understanding of configuration options and their interactions [83], the selection of an optimal configuration (tuning) [70, 25, 67], the performance prediction of arbitrary configurations [5, 44, 72] or the automated specialization of configurable systems [93, 92]. These works measure the performance of several configurations (a sample) under specific settings to then build a performance model. Input data further challenges these use-cases, since both the software configuration and the input spaces should be handled. Inputs also question and can threaten the generalization of configuration knowledge e.g., a performance prediction model for a given input may well be meaningless and inaccurate for another input.

On the one hand, some works have been addressing the performance analysis of software systems [76, 12, 23, 27, 53, 86] depending on different input data (also called workloads or benchmarks), but all of them only considered a rather limited set of configurations. On the other hand, works and studies on configurable systems usually neglect input data (e.g., using a unique video for measuring the configurations of a video encoder). We aim to combine both dimensions by performing an in-depth, controlled study of several configurable systems to make it vary in the large, both in terms of configurations and inputs. In contrast to research papers mixing multiple factors of the executing environment [42, 98], we concentrate on inputs and software configurations only, which allow us to draw reliable conclusions regarding the specific impact of inputs on software variability.

In this paper, we conduct, to our best knowledge, the first in-depth empirical study that measures how inputs individually interact with software variability. To do so, we systematically explore the impact of inputs and configuration options on the performance properties of 8 software systems. This study reveals that inputs fed to software systems can indeed interact with their options in non-monotonous ways, thus significantly impacting their performance properties. This observation questions the applicability of performance predictive models trained on only one input: are they still useful for other inputs? We then survey state-of-the-art papers on configurable systems to systematically explore the impact of inputs and configurations, but all of them only considered a rather limited set of configurations. To our best knowledge, the issue. Our contributions are as follows:

- The quantification of input sensitivity through several indicators and metrics per system and performance property;
- An analysis of how 65 state-of-the-art research papers on configurable systems address this problem in practice;
- A discussion on the impacts of our study (including key insights and open problems) for different engineering tasks of configurable systems (tuning, prediction, understanding, testing, etc.);
- Open science: a replication bundle that contains docker images, produced datasets of measurements and code.\footnote{Available on Github: https://github.com/llesoil/input_sensitivity/tree/master/}

The remainder of this paper is organized as follows: Section 2 explains the problem of input sensitivity and the research questions addressed in this paper. Section 3 presents the experimental protocol. Section 4 details the results. Section 5 shows how researchers address input sensitivity. Section 6 discusses the implications of our work. Section 7 details threats to validity. Section 9 summarizes key insights of our paper.

Typographic Convention. For this paper, we adopt the following typographic convention: emphasized will be relative to a software system, underlined to its configuration options and slanted to its performance properties.

2. Problem Statement

2.1. Sensitivity to Inputs of Configurable Systems

Configuration options of software systems can have different effects on performance (e.g., runtime), but so can the input data. For example, a configurable video encoder like x264 can process many kinds of inputs (videos) in addition to offering options on how to encode. Our hypothesis is that there is an interplay between configuration options and input data: some (combinations of) options may have different effects on performance depending on input.

Researchers observed input sensitivity in multiple fields, such as SAT solvers [105, 22], compilation [74, 16], video encoding [61], data compression [47]. However, existing studies either consider a limited set of configurations (e.g., only default configurations), a limited set of performance properties, or a limited set of inputs [1, 76, 12, 23, 27, 53, 86]. It limits some key insights about the input sensitivity of configurable systems. Valov et al. [98] studied the impact of hardware on software configurations, but fixed the input fed to software systems. Jamshidi et al. [41] explored how environmental conditions (hardware, input, and software versions) impact performances of software configurations. Besides considering a limited set of inputs (e.g., 3 input videos for x264), their study did not aim to isolate the individual effects of
input data on software configurations. As a result, it is impossible to draw reliable conclusions about the specific variability factors – among hardware, inputs and versions.

This work details, to the best of our knowledge, the first systematic empirical study that analyzes the interactions between input data and configuration options for different configurable systems. Through four research questions introduced in the next section, we characterize the input sensitivity problem and explore how this can alter our understanding of software variability.

2.2. Research Questions

When a developer provides a default configuration for its software system, one should ensure it will perform at best for a large panel of inputs. That is, this configuration will be near-optimal whatever the input. Hence, a hidden assumption is that two performance distributions over two different inputs are somehow related and close. In its simplest form, there could be a linear relationship between these two distributions: they simply increase or decrease with each other. **RQ1 - To what extent are the performance distributions of configurable systems changing with input data?** To answer this, we compute and compare performance distributions of different inputs. For software systems, unstable performance distributions across inputs induce that their optimal configuration change with their inputs. In particular, the default configuration should be adapted according to their input data.

But configuration options influence software performance, e.g., the energy consumption [13]. An option is called influential for a performance when its values have a strong effect on this performance [42, 18]. For example, developers might wonder whether the option they add to a configurable system has an influence on its performance. However, is an option identified as influential for some inputs still influential for other inputs? If not, it would become both tedious and time-consuming to find influential options on a per-input basis. Besides, it is unclear whether activating an option is always worth it in terms of performance; an option could improve the overall performance while reducing it for few inputs. If so, users may wonder which options to enable to improve software performance based on their input data. **RQ2 - To what extent the effects of configuration options are consistent with input data?** In this question, we quantify how the effects and importance of software options change with input data. If this change is significant, tuning these options to optimize performance should be adapted to the current input.

**RQ1** and **RQ2** study how inputs affect (1) performance distributions and (2) the effects of different configuration options. However, the performance distributions could change in a negligible way, without affecting the software user’s experience. Before concluding on the real impact of the input sensitivity, it is necessary to quantify how much this performance changes from one input to another. **RQ3 - How much performance are lost when reusing a configuration across inputs?** In particular, we estimate the loss in performance when configuring a software while ignoring the input sensitivity to inputs. To put it more positively, this loss is also the potential gain, in terms of performance, to tune a software system for its input data. **RQ1** to **RQ3** present the problem of input sensitivity. The fourth and last question explores the limits of this problem and give insights on how to address it concretely. Though all inputs are different, the number of possible interactions between the software systems and the processed inputs is limited. Therefore, there might exist inputs interacting in the same way with the software, and thus having similar performance profiles. **RQ4 - What is the benefit of grouping the inputs?** For this question, we form, analyze and characterize different groups of inputs having similar performance distributions and show the benefits of these groups to address the input sensitivity issue.

3. Experimental protocol

We designed the following experimental protocol to answer these research questions.

**3.1. Data Collection**

We first collect measurements of systems processing inputs.

**Protocol.** Figure 2 depicts the step-by-step protocol we respect to measure performance of software systems. Each line of Table 1 should be read following Figure 2: System with Steps 1 and 2; Configurations #C with Step 3; the nature of inputs I and their number #I with Step 4; Performance P with Steps 5 and 6; Docker links a container for executing all the steps and Data the datasets containing the performance measurements. Figure 2 shows in beige an example with the x264 encoder. Hereafter, we provide details for each step of the protocol.
Steps 1 & 2 - Software Systems. We consider 8 software systems. We choose them because they are open-source, well-known in various fields and already studied in the literature: gcc [74, 10], the compiler for gnu operating system; ImageMagick [88, 87], a software system processing pictures and images; lingeling [36, 22], a SAT solver; node.js [38, 89], a widely-used JavaScript execution environment; poppler [48, 58], a library designed to process .pdf files; SQLite [92, 41], a database manager system; x264 [41, 5], a video encoder based on H264 specifications; xz [98, 64], a file system manager. We also choose these systems because they handle different types of input data, allowing us to draw conclusions as general as possible. For each software system, we use a unique private server with the same configuration running over the same operating system.2 We download and compile a unique version of the binary; for PDF rendering, we download and compile a unique version of lingeling, which should use stronger --debloc settings.3 Out of these configuration options, we then sample #C configurations by using random sampling [73]. In the previous example, after the selection of --mbtree and --debloc, the sampling step would generate multiple configurations with combinations of options’ values: C1, with --mbtree activated and --debloc set to "0:0"; C2, with --mbtree deactivated and --debloc set to "-2:-2"; C3, with --mbtree deactivated and --debloc set to "0:0". To ensure that each value of a software option is well represented in the final set of configurations, we statistically test the uniformity of its values. To do so, we apply a Kolmogorov-Smirnov test [60] to each value of our eight software systems.4 In the previous example, for a boolean option like --mbtree that can be either activated or deactivated, a valid Kolmogorov-Smirnov test guarantees that --mbtree is activated in roughly 50% of the configurations. To mitigate the threat of only using random sampling, we also considered various informed configurations picked in the documentation. For instance, for x264, we considered the ten presets configurations recommended by the documentation5

Step 4 - Inputs I. For each system, we select a different set of input data: for gcc, PolyBench v3.1 [78]; for ImageMagick, a sample of ImageNet [14] images (from 1.1 kB to 7.3 MB); for lingeling, the 2018 SAT competition’s benchmark [36]; for node.js, its test suite; for poppler, the Trent Nelson’s PDF Collection [68]; for SQLite, a set of generated TPC-H [75] databases (from 10 MB to 6 GB); for x264, the YouTube User General Content dataset [101] of videos (from 2.7 MB to 39.7 GB); for xz, a combination of the Silesia [15] and the Canterbury [8] corpus. We choose them because these are large and freely available datasets of inputs, well-known in their field and already used by researchers and practitioners.

Steps 5 & 6 - Performance properties P. For each system, we systematically execute all the configurations of C on all the inputs of I. For the #M resulting executions, we measure as many performance properties as possible: for gcc, ctime and exec the times needed to compile and execute a program and the size of the binary; for ImageMagick, the time to apply a Gaussian blur [37] to an image and the size of the resulting image; for lingeling, the number of conflicts and reductions found in 10 seconds of execution; for node.js, the number of operations per second (ops) executed by the script; for poppler, the time needed to extract the images of the pdf, and the size of the images; for SQLite, the time needed to answer 15 different queries q1 \rightarrow q15; for x264, the bitrate (the average amount of data encoded per second), the cpu usage (percentage), the average number of frames encoded per second (fps), the size of the compressed video and the elapsed time; for xz, the size of the compressed file, and the time needed to compress it. It results in a set a tabular data, one for each input and each software system, consisting of a list of configurations with their performance property values.

Replication. To allow researchers to easily replicate the measurement process, we provide a docker container for each system (see the links in the Docker column of

<table>
<thead>
<tr>
<th>System</th>
<th>Domain</th>
<th>Commit</th>
<th>#LoCs</th>
<th>Configs</th>
<th>#C</th>
<th>Inputs I</th>
<th>#I</th>
<th>#M</th>
<th>Performance(s) P</th>
<th>Docker</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>gcc</td>
<td>Compilation</td>
<td>ce4d607</td>
<td>9066</td>
<td>697</td>
<td>80</td>
<td>.c programs</td>
<td>30</td>
<td>2400</td>
<td>size, ctime, exec</td>
<td>Link</td>
<td>Link</td>
</tr>
<tr>
<td>ImageMagick</td>
<td>Image processing</td>
<td>80e45b6</td>
<td>6489</td>
<td>964</td>
<td>100</td>
<td>images</td>
<td>1000</td>
<td>100000</td>
<td>size, time</td>
<td>Link</td>
<td>Link</td>
</tr>
<tr>
<td>lingeling</td>
<td>SAT solver</td>
<td>745f772</td>
<td>33402</td>
<td>100</td>
<td>100</td>
<td>SAT formulae</td>
<td>351</td>
<td>35100</td>
<td>#conf, #red.</td>
<td>Link</td>
<td>Link</td>
</tr>
<tr>
<td>nodeJS</td>
<td>JS runtime env.</td>
<td>7f834b4b</td>
<td>6205</td>
<td>618</td>
<td>50</td>
<td>.js scripts</td>
<td>1939</td>
<td>96950</td>
<td>#operations/s</td>
<td>Link</td>
<td>Link</td>
</tr>
<tr>
<td>poppler</td>
<td>PDF rendering</td>
<td>42d56088</td>
<td>1970</td>
<td>19</td>
<td>16</td>
<td>.pdf files</td>
<td>1480</td>
<td>23680</td>
<td>size, time</td>
<td>Link</td>
<td>Link</td>
</tr>
<tr>
<td>SQLite</td>
<td>DBMS</td>
<td>5a30525</td>
<td>32049</td>
<td>50</td>
<td>150 databases</td>
<td>7500</td>
<td>15 query times q1-q15</td>
<td>Link</td>
<td>Link</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x264</td>
<td>Video encoding</td>
<td>e9a5590</td>
<td>11064</td>
<td>201</td>
<td>201</td>
<td>videos</td>
<td>1397</td>
<td>280797</td>
<td>cpu, fps, kbs, size, time</td>
<td>Link</td>
<td>Link</td>
</tr>
<tr>
<td>xz</td>
<td>Data compression</td>
<td>87da44d</td>
<td>37489</td>
<td>30</td>
<td>48</td>
<td>system files</td>
<td>1440</td>
<td>14400</td>
<td>size, time</td>
<td>Link</td>
<td>Link</td>
</tr>
</tbody>
</table>

2 The configurations of the running environments are available at: https://github.com/llesoil/input_sensitivity/tree/master/replication/Environments.md

3 See the documentation of x264 at https://silentaperture.github.io/mbtree-guide/encoding/x264.html

4 See the documentation of ImageMagick at https://github.com/llesoil/input_sensitivity/tree/master/results/others/configs/sampling.md

5 See http://www.chanseru.com/Boku/HLs/x264_Settings.html#preset
3.2. Performance Correlations (RQ1)

Based on the analysis of the data collected in Section 3.1, we can now answer the first research question: RQ1 - To what extent are the performance distributions of configurable systems changing with input data? To check this hypothesis, we compute, analyze and compare the Spearman’s rank-order correlation [46] of each couple of inputs for each system. It is appropriate in our case since all performance properties are quantitative variables measured on the same set of configurations.

Spearman correlations. The correlations are considered as a measure of similarity between the configurations’ performance over two inputs. We compute the related p-values: a correlation whose p-value is higher than the chosen threshold 0.05 is considered as null. We use the Evans rule [21] to interpret these correlations. In absolute value, we refer to correlations by the following labels: very low: 0-0.19, low: 0.2-0.39, moderate: 0.4-0.59, strong: 0.6-0.79, very strong: 0.8-1.00. A negative score tends to reverse the ranking of configurations. Very low or negative scores have practical implications: a good configuration for an input can very well exhibit bad performance for another input.

3.3. Effects of Options (RQ2)

To understand how a performance model can change based on a given input, we next study how input data interact with configuration options. RQ2 - To what extent the effects of configuration options are consistent with input data? To assess the relative significance and effect of options, we use two well-known statistical methods [81, 9], also widely used in the context of interpretable machine learning and configurable systems [72, 62, 41]. For instance, Jamshidi et al. [41] used similar indicators to measure the sensitivity of configurations regarding computing environmental conditions (hardware, input, and software versions).

Random forest importances. The tree structure provides insights about the most essential options for prediction, because such a tree first splits w.r.t. options that provide the highest information gain. We use random forests [9], a vote between multiple decision trees: we can derive, from the forests trained on the inputs, estimates of the options importance. The computation of option importance is realized through the observation of the effect on random forest accuracy when randomly shuffling each predictor variable [62]. For a random forest, we consider that an option is influential if the median (on all inputs) of its option importance is greater than \( \frac{1}{n_{\text{opt}}} \) where \( n_{\text{opt}} \) is the number of options considered in the dataset. This threshold represents the theoretic importance of options for a software having equally important options.

Linear regression coefficients. The coefficients of an ordinary least square regression [81] weight the effect of configuration options. These coefficients can be positive (resp. negative) if a bigger (resp. lower) option value results in a bigger performance. Ideally, the sign of the coefficients of a given option should remain the same for all inputs: it would suggest that the effect of an option onto performance is stable. We also provide details about coefficients related to feature interactions [96, 80] in the companion repository.

3.4. Impact of Inputs on Performance (RQ3)

To complete this experimental protocol, we ask whether adapting the software to its input data is worth the cost of finding the right set of parameters i.e., the concrete impact of input sensitivity. RQ3 - How much performance are lost when reusing a configuration across inputs? To estimate how much we can lose, we first define two scenarios of reuse \( S_1 \) and \( S_2 \):

- **S1 - Baseline.** In this scenario, we value input sensitivity and just train a simple performance model on a target input. We choose the best configuration according to the model, configure the related software with it and execute it on the target input.
- **S2 - Ignoring input sensitivity.** In this scenario, let us pretend that we ignore the input sensitivity issue. We train a model related to a given input i.e., the source input, and then predict the best configuration for this source input. If we ignore the issue of input sensitivity, we should be able to easily reuse this model for any other input, including the target input of \( S_1 \). Finally, we execute the software with the predicted configuration on the target input.

In this part, we systematically compare \( S_1 \) and \( S_2 \) in terms of performance for all inputs, all performance properties and all software systems. For \( S_1 \), we repeat the scenario ten times with different sources, uniformly chosen among other inputs and compute the average performance. For both scenarios, due to the imprecision of the learning procedure, the models can recommend sub-optimal configurations. Since this imprecision can alter the results, we consider an ideal case for both scenarios and assume that the performance models always recommend the best possible configuration.

Performance ratio. To compare \( S_1 \) and \( S_2 \), we use a performance ratio i.e., the performance obtained in \( S_1 \) over the performance obtained in \( S_2 \). If the ratio is equal to 1, there is no difference between \( S_1 \) and \( S_2 \) and the input sensitivity does not exist. A ratio of 1.4 would suggest that the performance of \( S_1 \) is worth 1.4 times the performance of \( S_2 \); therefore, it is possible to gain up to \((1.4 - 1) \times 100 = 40\%\) performance by choosing \( S_1 \) instead of \( S_2 \). We also report on the standard deviation of the performance.

---

6 Guidelines for replication are available at: https://github.com/llesoil/input_sensitivity/tree/master/replication/README.md
3.5. Groups of Inputs (RQ4)

Lastly, we explore how the issue of input sensitivity can be concretely addressed. For mitigating input sensitivity, an idea is to group together inputs based on their performance distributions. \textbf{RQ4 - What is the benefit of grouping the inputs?} The inputs belonging to the same group are supposed to share common properties and be processed in a similar manner by the software [16]. We perform hierarchical clustering [69] to gather inputs having similar performance profiles.

\textbf{Hierarchical clustering.} This technique considers the correlations between performance distributions as a measure of similarity between inputs. Based on these values, it forms groups of inputs minimizing the intra-class variance (discrepancy of performance among a group) and maximizing the inter-class variance (discrepancy of performance between different groups of inputs). As linkage criteria, we choose the Ward method [69] since in our case, (1) single link (minimum of distance) leads to numerous tiny groups (2) centroid or average tend to split homogeneous groups of inputs and (3) complete link aggregates unbalanced groups. As a metric, we kept the Euclidian distance - used as default. We manually select the final number of groups.

For each group, we then report on few key indicators summarizing the specifics of inputs’ performance: the Spearman correlations between performance distributions of inputs (RQ1), the importance and effects of options (RQ2) as well as few properties characterizing the inputs e.g., the spatial complexity of an input video or the number of lines of a .c program. We compare their average value in the different groups.

4. Results

4.1. Performance Correlations (RQ1)

We first explain the results of RQ1 and their consequences on the poppler use case i.e., an extreme case of input sensitivity, and then generalize to our other software systems.

\textbf{Extract images of input pdfs with poppler.} The content of pdf files fed to poppler may vary; the input pdf can contain a 2-page extended abstract with plain text, a 10-page conference article with few figures or a 300-page book full of pictures. Depending on this content, extracting the images embedded in those files can be quicker or slower for the same configuration. Moreover, different configurations could be adapted for the conference paper but not for the book (or conversely), leading to different rankings of extraction time and thus different rank-based correlation values.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{Spearman rank based correlations - poppler, \texttt{time}.}
\end{figure}

Figure 3 depicts the Spearman rank-order correlations of extraction time between pairs of input pdfs fed to poppler. Results suggest a positive correlation (see dark red cells), though there are pairs of inputs with lower (see white cells) and even negative (see dark blue cells) correlations. More than a quarter of the correlations between input pdfs are positive and at least moderate - third quartile Q3 greater than 0.52.

\textbf{Meta-analysis.} Over the 8 systems,\textsuperscript{7} we observe different cases. There exist software systems not sensitive at all to inputs. In our experiment, gcc, imagemagick and xz present almost exclusively high and positive correlations between inputs e.g., Q1 = 0.82 for the compressed size and xz. For these, un- or negatively-correlated inputs are an exception more than a rule. In contrast, there are software systems, namely lingeling, nodeJS, SQLite and poppler, for which performance distributions completely change and depend on input data e.g., Q2 = 0.09 for nodeJS and ops, Q3 = 0.12 for lingeling and conflicts. For these, we draw similar conclusions as in the poppler case. In between, x264 is only input-sensitive w.r.t. a performance property; it is for bitrate and size but not for cpu, fps and time e.g., 0.29 as deviation for size against 0.08 for time.

\textsuperscript{7}Detailed RQ1 results for other systems are available at: https://github.com/llesoil/input_sensitivity/tree/master/results/RQS/RQ1/RQ1.md

ratio distribution. A standard deviation of 0 implies that we gain or lose the same proportion of performance when picking $S_1$ over $S_2$. 

The inputs belonging to the same group are supposed to share common properties and be processed in a similar manner by the software [16]. We perform hierarchical clustering [69] to gather inputs having similar performance profiles.

\textbf{Hierarchical clustering.} This technique considers the correlations between performance distributions as a measure of similarity between inputs. Based on these values, it forms groups of inputs minimizing the intra-class variance (discrepancy of performance among a group) and maximizing the inter-class variance (discrepancy of performance between different groups of inputs). As linkage criteria, we choose the Ward method [69] since in our case, (1) single link (minimum of distance) leads to numerous tiny groups (2) centroid or average tend to split homogeneous groups of inputs and (3) complete link aggregates unbalanced groups. As a metric, we kept the Euclidian distance - used as default. We manually select the final number of groups.

For each group, we then report on few key indicators summarizing the specifics of inputs’ performance: the Spearman correlations between performance distributions of inputs (RQ1), the importance and effects of options (RQ2) as well as few properties characterizing the inputs e.g., the spatial complexity of an input video or the number of lines of a .c program. We compare their average value in the different groups.

4. Results

4.1. Performance Correlations (RQ1)

We first explain the results of RQ1 and their consequences on the poppler use case i.e., an extreme case of input sensitivity, and then generalize to our other software systems.

\textbf{Extract images of input pdfs with poppler.} The content of pdf files fed to poppler may vary; the input pdf can contain a 2-page extended abstract with plain text, a 10-page conference article with few figures or a 300-page book full of pictures. Depending on this content, extracting the images embedded in those files can be quicker or slower for the same configuration. Moreover, different configurations could be adapted for the conference paper but not for the book (or conversely), leading to different rankings of extraction time and thus different rank-based correlation values.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{Spearman rank based correlations - poppler, \texttt{time}.}
\end{figure}

Figure 3 depicts the Spearman rank-order correlations of extraction time between pairs of input pdfs fed to poppler. Results suggest a positive correlation (see dark red cells), though there are pairs of inputs with lower (see white cells) and even negative (see dark blue cells) correlations. More than a quarter of the correlations between input pdfs are positive and at least moderate - third quartile Q3 greater than 0.52.

\textbf{Meta-analysis.} Over the 8 systems,\textsuperscript{7} we observe different cases. There exist software systems not sensitive at all to inputs. In our experiment, gcc, imagemagick and xz present almost exclusively high and positive correlations between inputs e.g., Q1 = 0.82 for the compressed size and xz. For these, un- or negatively-correlated inputs are an exception more than a rule. In contrast, there are software systems, namely lingeling, nodeJS, SQLite and poppler, for which performance distributions completely change and depend on input data e.g., Q2 = 0.09 for nodeJS and ops, Q3 = 0.12 for lingeling and conflicts. For these, we draw similar conclusions as in the poppler case. In between, x264 is only input-sensitive w.r.t. a performance property; it is for bitrate and size but not for cpu, fps and time e.g., 0.29 as deviation for size against 0.08 for time.

\textsuperscript{7}Detailed RQ1 results for other systems are available at: https://github.com/llesoil/input_sensitivity/tree/master/results/RQS/RQ1/RQ1.md


RQ1 - To what extent are the performance distributions of configurable systems changing with input data? We show that: (1) depending on the inputs, the rank-based correlations of performance distribution can be high, close to zero, or even negative; (2) since configuration rankings can change with input data, the best configuration for an input will not be the best configuration for another input. The consequence is that one cannot blindly reuse a configuration prediction model across inputs and that developers should not provide to end-users a unique default configuration whatever the input is.

4.2. Effects of Options (RQ2)

We first explain the results of RQ2 and their concrete consequences on the bitrate of x264 - an input-sensitive case, to then generalize to other software systems.

**Encoding input videos with x264.** Figures 4a and 4b report on respectively the boxplots of configuration options’ feature importances and effects when predicting x264’s bitrate for all input videos. On the top graph, we displayed the boxplots of the distribution of importances for each option (y-axis). On the bottom graph, we displayed the boxplots of the distribution of regression coefficients for each option (y-axis). Each red square is representing a model trained on one input, and all of them constitute the resulting distribution. Each algorithm is using 100% of the configurations in the training set. We also compute the Mean Absolute Percentage Error (MAPE) for all the systems, inputs and non-functional properties when predicting the performance with random forests and linear regression. For random forest, we can ensure that our models are giving a good prediction, the median value of the MAPE across all inputs being systematically under 5% for each couple of software systems and performance properties. For Linear Regression, results tend to show higher values of MAPE, suggesting that the configurations spaces are too hard to learn from for such simple models. Other variants of feature importances and linear regression (permutation importance,9 drop-column importance10 and Shapley values11) have been computed to ensure the robustness of results in the companion repository. They reached similar results, which confirms our conclusions with the chosen indicators.

Three options are strongly influential for a majority of videos on Figure 4a: subme, mbtree and aq-mode, but their importance can differ depending on input videos: for instance, the importance of subme is 0.83 for video #1365 and only 0.01 for video #40. Because influential options vary with inputs, performance models and approaches based on feature selection [62] such as performance-influence model [83, 102] may not generalize well to all input videos.

Most of the options have positive and negative coefficients on Figure 4b; thus, the specific effects of options heavily depend on input videos. It is also true for influential options: mbtree can have positive and negative (influential) effects on the bitrate i.e., activating mbtree may be worth only for few input videos. The consequence is that tuning the options of a software system should be adapted to the current input, and not done once for all the inputs.

**Meta-analysis.** For gcc, imagemagick and xz, the importances are quite stable. As an extreme case of stability, the importances of the compressed size for xz are exactly the same, except for two inputs. For these systems, the coefficients of linear regression mostly keep the same sign across inputs i.e., the effects of options do not change with inputs. For input-sensitive software systems, we always observe high variations of options’ effects (lingeling,
poppler or SQLite), sometimes coupled to high variations of options’ importances (nodeJS). For instance, the option format for poppler can have an importance of 0 or 1 depending on the input. For all software systems, there exists at least one performance property whose effects are not stable for all inputs e.g., one input with negative coefficient and another with a positive coefficient. For x264, it depends on the performance property; for cpu, fps and time, the effect of influential options are stable for all inputs, while for the bitrate and the size, we can draw the conclusions previously presented.

**RQ2 - To what extent the effects of configuration options are consistent with input data?**

Two lessons learned: (1) the importance of software options change with input data, implying that an option can be influential only for few input data, but not for the rest of the inputs; (2) the effect of software options on performance properties vary with input data. An option can have a positive influence for an input and at the same time a negative influence for another input. As a result, tuning the options of a software system should depend on its processed inputs.

**4.3. Impact of Inputs on Performance (RQ3)**

This section presents the evaluation of RQ1,2 w.r.t. the protocol of Section 3.4. Figure 5 presents the loss of performance (y-axis, in %) due to input sensitivity for the different software systems and their performance properties.

![Figure 5: Performance loss (%) when ignoring input sensitivity per system and performance property (x-axis)](image)

**Key result.** The average performance ratio across all the software systems is 1.38: we can expect an average drop of 38% in terms of performance when ignoring the input sensitivity.3

**Meta-analysis.** For software systems whose performance are stable across inputs (gcc, imagemagick and xz), there are few differences between inputs. For instance, for the output size of xz, there is no variation between scenarios S1 (i.e., using the best configuration) and S2 (i.e., reusing a the best configuration of a given input for another input): all performance ratios (i.e., performance S1 over performance S2) are equals to 1 whatever the input.

For input-sensitive software systems (lingeling, nodeJS, SQLite and poppler), changing the configuration can lead to a negligible change in a few cases. For instance, for the time to answer the first query q1 with SQLite, the median is 1.03; in this case, SQLite is sensitive to inputs, but its variations of performance -less than 4%– do not justify the complexity of tuning the software. But it can also be a huge change; for lingeling and solved conflicts, the 95th percentile ratio is equal to 8.05 i.e., a factor of 8 between S1 and S2. It goes up to a ratio of 10.11 for poppler’s extraction time; there exists an input pdf for which extracting its images is ten times slower when reusing a configuration compared to the fastest.

In between, x264 is a complex case. For its low input-sensitive performance (e.g., cpu and etime), it moderately impacts the performance when reusing a configuration from one input to another - average ratios at resp. 1.42 and 1.43. In this case, the rankings of performance do not change a lot with inputs, but a small ranking change does make the difference in terms of performance.

On the contrary, for the input-sensitive performance (e.g., the bitrate), there are few variations of performance: we can lose $1 - \frac{1}{11} \approx 9\%$ of bitrate in average. In this case, it is up to the compression experts to decide; if losing up to $1 - \frac{1}{11} \approx 24\%$ of bitrate is acceptable, then we can ignore input sensitivity. Otherwise, we should consider tuning x264 for its input video.

**RQ3 - How much performance are lost when reusing a configuration across inputs?** In average, randomly reusing configurations across inputs leads to a performance drop of 38%, which suggests we cannot ignore input sensitivity. On the good side, performance can be multiplied up to a ratio of 10 if we tune other systems for their input data.

**4.4. Groups of Inputs (RQ4)**

We illustrate the results of this section using the bitrate of x264 when encoding input videos.4 In Figure 6, we first compute the correlations between performance of all input videos, as in RQ1. Then, we perform hierarchical clustering on x264 measurements to gather inputs having similar bitrate distributions and visually group correlated videos together. The resulting groups are delimited and numbered directly in the figure. For instance, the group (1) is located in the top-left part of the correlogram by the triangle (1).

---

12Detailed RQ3 results for other performance properties are available at: [https://github.com/llesoil/input_sensitivity/tree/master/results/RQ1/RQ3.md](https://github.com/llesoil/input_sensitivity/tree/master/results/RQ1/RQ3.md)

13To compute this result, we removed SQLite biasing the results with its 15 performance properties.

14The results for the rest of software systems can be consulted in the companion repository at [https://github.com/llesoil/input_sensitivity/blob/master/results/RQ2/RQ4/groups.ipynb](https://github.com/llesoil/input_sensitivity/blob/master/results/RQ2/RQ4/groups.ipynb)
Group description. In total, we isolate four groups of input videos. These groups are presented and described in Table 2:

- Group 1 is mostly composed of moving or action videos, often picked in the sports or news categories and with high spatial and chunk complexities;
- Group 2 gathers large input videos, with big resolution videos, taken for instance in the High Dynamic Range category. They typically have a low spatial complexity and a high temporal complexity;
- Group 3 is composed of "still image" videos i.e., input videos with few changes of background, with low temporal and chunk complexities. A typical example of this kind of video would be a course with a fixed board, chosen in the Lecture or in the HowTo category;
- Group 4 is a group of average videos with average property values and various contents.

Revisiting RQ1. In a group of inputs, performance distributions of inputs are highly correlated with each other - positively, strong or very strong. The input videos of the same group have similar bitrate rankings; their performance react the same way to the same configurations of x264. However, the group 1 is uncorrelated (very low, low) or negatively correlated (moderate, strong and very strong) with the group 2 - see the intersection area between triangles 1 and 2. In this case, a single configuration of x264 working for the group 1 should not be reused directly on a video of the group 2. So, these groups are capturing the difference of performance between inputs; once in a group, input sensitivity does not represent a problem anymore.

Revisiting RQ2. Within a group, the effect and importance of options are stable and the inputs all react the same way to the same options, while they differ between the different groups. For instance, for the group 1, aq-mode is influential (Imp = 0.27), while it is not for the group 3 (Imp = 0.04). Likewise, the effects of mbtree vary with the group of inputs; for the group 1, activating mbtree always increases the bitrate (Effect = +0.33), while for the groups 2, 3 and 4, it diminishes the bitrate (Effects = -0.68, -0.42, -0.11 respectively). Under these circumstances, configuring the software system once per group of inputs is probably a reasonable solution for tackling input sensitivity.

Revisiting RQ3. For the bitrate of x264, reusing a configuration from a source input to a target input generates a lower performance drop if the source and the target inputs are selected in the same group (e.g., 13% for group 2) compared to a random selection (34% in general). If we are able to find the best configuration for one input video in a group, this configuration will be good-enough for the rest of the inputs in this group.

Meta-analysis. For the other input-sensitive systems, results tend to show similar results as for x264 and the bitrate. For instance, with poppler, grouping tend to gather inputs with the same influence and effect of options; for the size, the importance of format is influential in groups 2 and 4 but not in groups 1 and 3; for the execution time, in groups 1 and 2, -jp2 has a positive effect overall while a negative effect for groups 3 and 4. When the groups discriminate inputs with different effects of options, RQ3 results tend to be more impressive e.g., the average performance loss of 49% vanishes when grouping the inputs (in the four groups, 1%-4%-12%-2% when reusing a configuration inside the group). The same applies for Nodejs: --jitless is influential for groups 2, 3 and 4 but not for group 1. In RQ3 results, the average performance drop of 44% becomes 7.4%-6.7%-13.7%-9.0% in the four groups. For non input-sensitive systems, we do not observe such difference between the groups. Grouping seems to be ineffective; the same effect of options are observed; it does not change the performance loss, already low e.g., for the execution time of imagemagick, 6% in general and 1%-1%-2%-6% in the groups.
Classify inputs into groups. In short, grouping together inputs seems a right approach to reduce input sensitivity. However there is now a problem: we need to map a given input into a group \textit{a priori}, without having access to all measurements. Since those four groups are consistent and share common properties, one domain expert or one machine learning model could classify these inputs \textit{a priori} into a group without measuring their performance just by looking at the properties of the inputs.

Benefits for benchmarking. These groups allow to increase the representativeness of profiles of inputs used to test software systems, while greatly lowering the number of inputs of this set. In the companion repository, we operate on previous results to create a short but representative set of input videos dedicated to the benchmarking of x264: we reduce the dataset, initially composed of 1397 input videos [101], to a subset of 8 videos, selecting 2 cheap videos in each group of performance.\footnote{See the resulting benchmark and its construction at: \url{https://github.com/llesoil/input_sensitivity/tree/master/results/RQS/RQ4/x264_bitrate.md}}

RQ4 - What is the benefit of grouping the inputs? Grouping inputs together is beneficial to apply in input-sensitive systems: the performance distributions, the influence of options, and the effect of options are alike between inputs of the same group. To classify inputs in one of those groups, the characteristics of inputs can be used without measuring any configuration. These groups can also be derived to create short but representative sets of inputs designed to benchmark software systems. For other systems, it does not bring any significant improvements.

5. Sensitivity to Inputs in Research

In this section, we explore the significance of the input sensitivity problem in research. Do researchers know the issue of input sensitivity? How do they deal with inputs in their papers? Is the interaction between software configurations and input sensitivity a well-known issue?

5.1. Experimental Protocol

First, we aim at gathering research papers \cite{29} predicting the performance of configurable systems \textit{i.e.}, with a performance model \cite{30}.

Gather research papers. We focused on the publications of the last ten years. To do so, we analyzed the papers published (strictly) after 2011 from the survey of Pereira \textit{et al.} \cite{72} - published in 2019. We completed those papers with more recent papers (2019-2021), following the same procedure as in \cite{72}. We have only kept research work that trained performance models.

Search for input sensitivity. We read each selected paper and answered four different questions: Q-A. Is there a software system processing input data in the study? If not, the impact of input sensitivity in the existing research work would be relatively low. The idea of this research question is to estimate the proportion of the performance models that could be affected by input sensitivity. Q-B. Does the experimental protocol include several inputs? If not, it would suggest that the performance model only captures a partial truth, and might not generalize for other inputs fed to the software system. Q-C. Is the problem of input sensitivity mentioned in the paper? This question aims to state whether researchers are aware of the input sensitivity issue, and estimate the proportion of the papers that mention it as a potential threat to validity. Q-D. Does the paper propose a solution to generalize the performance model across inputs? Finally, we check whether the paper proposes a solution managing input sensitivity \textit{i.e.}, if the proposed approach could be adapted to our problem and predict a near-optimal configuration for any input. The results were obtained by one author and validated by all other co-authors.

5.2. How do Research Papers Address Input Sensitivity?

Table 3 lists the 65 research papers we identified following this protocol, as well as their individual answers to Q-A→Q-D. A checked cell indicates that the answer to the corresponding question (column) for the corresponding paper (line) is yes. Since answering Q-B, Q-C or Q-D only makes sense if Q-A is checked, we grayed and did not consider Q-B, Q-C and Q-D if the answer of Q-A is no. We also provide full references and detailed justifications in the companion repository.\footnote{List of papers at \url{https://github.com/llesoil/input_sensitivity/tree/master/results/RQS/RQ6/}} We now comment the average results:

Q-A. Is there a software system processing input data in the study? Of the 65 papers, 60 (94\%) consider at least one configurable system processing inputs. This large proportion gives credits to input sensitivity and its potential impact on research work.

Q-B. Does the experimental protocol include several inputs? 63\% of the research work answering yes to Q-A include different inputs in their protocol. But what about the other 37\%? It is understandable not to consider several inputs because of the cost of measurements. However, if we reproduce all experiments of Table 3 using other input data, will we draw the same conclusions for each paper? Based on the results of RQ1 → RQ3, we encourage researchers to consider at least a set of inputs in their protocol (see Section 6).

Q-C. Is the problem of input sensitivity mentioned in the paper? Only half (47\%) of the papers mention the issue of input sensitivity, mostly without naming it or using a domain-specific keyword \textit{e.g.}, workload variation \cite{98}. For the other half, we cannot guarantee with certainty that input sensitivity concerns all papers. But we shed light on...
Table 3: Input sensitivity in research. Q-A. Is there a software system processing input data in the study? Q-B. Does the experimental protocol include several inputs? Q-C. Is the problem of input sensitivity mentioned in the paper? Q-D. Does the paper propose a solution to generalize the performance model across inputs? Justifications in the companion repository.
this issue: ignoring input sensitivity can prevent the generalization of performance models across inputs. This is especially true for the 37% of papers answering no to Q-B i.e., considering one input per system: only 14% of these research works mention it.

**Q-D.** Does the paper propose a solution to generalize the performance model across inputs? We identified 15 papers [99, 16, 20, 77, 41, 98, 42, 51, 97, 52, 39, 63, 56, 57, 17] proposing contributions that may help in better managing the input sensitivity problem. However, most of them have not been designed to operate over actual inputs, but rather changes of computing environments [41, 98, 42, 51, 97, 52, 39]. Other works [99, 16, 20, 77, 57] only apply it to a specific domain (database [99], compilation [16], cloud computing [20, 57, 17] or programs analysis [77]) with open questions about applicability and effectiveness in other areas. We plan to confront these techniques on our dataset for multiple systems.

**Conclusion.** While half of the research articles mention input sensitivity, few actually address it, and most often on a single system and domain. Input sensitivity can affect multiple research works and questions their practical relevance for a field deployment.

### 6. Implications of our study

In this section, we first summarize results of our study and then discuss their impacts on several research directions.

#### Table 4: Summary of the input sensitivity on our dataset

<table>
<thead>
<tr>
<th>System</th>
<th>Perf.</th>
<th>Correlation $[C_{min}, C_{max}]$</th>
<th>Most infl. opt. Effect [min, max]</th>
<th>Impact (%)</th>
<th>IS Score</th>
<th>RQ</th>
<th>RQ</th>
<th>RQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>gcc</td>
<td>ctime</td>
<td>[0.72, 0.97]</td>
<td>[−0.26, −0.18]</td>
<td>13</td>
<td>0.32</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>exec</td>
<td>[−0.69, 1]</td>
<td>[−0.21, 0.64]</td>
<td>27</td>
<td>0.92</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>image</td>
<td>time</td>
<td>[−0.24, 1]</td>
<td>[−0.18, 0.95]</td>
<td>4</td>
<td>0.39</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>lingeling</td>
<td># conf</td>
<td>[−0.9, 0.92]</td>
<td>[−0.79, 0.91]</td>
<td>15</td>
<td>0.75</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td># redag</td>
<td>[−0.99, 1]</td>
<td>[−0.79, 0.91]</td>
<td>10</td>
<td>0.7</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>nodejs</td>
<td>spce</td>
<td>[−0.87, 0.95]</td>
<td>[−1 &lt; 0.95]</td>
<td>17</td>
<td>0.79</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>poppler</td>
<td>time</td>
<td>[−0.94, 1]</td>
<td>[−0.94, 1]</td>
<td>37</td>
<td>0.98</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>SQLlite</td>
<td>q1</td>
<td>[−0.78, 0.87]</td>
<td>[−0.69, 0.58]</td>
<td>2</td>
<td>0.45</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>q15</td>
<td>[−0.3, 0.94]</td>
<td>[−0.59, 0.6]</td>
<td>3</td>
<td>0.37</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>x264</td>
<td>bitrate</td>
<td>[−0.69, 1]</td>
<td>[−0.55, 0.28]</td>
<td>21</td>
<td>0.84</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>cpu</td>
<td>[−0.31, 1]</td>
<td>[−0.1, 0.84]</td>
<td>7</td>
<td>0.47</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>fps</td>
<td>[0.01, 1]</td>
<td>[−0.62, 0.23]</td>
<td>6</td>
<td>0.37</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>size</td>
<td>[−0.69, 1]</td>
<td>[−0.58, 0.28]</td>
<td>21</td>
<td>0.84</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>time</td>
<td>[0.02, 1]</td>
<td>[−0.17, 0.45]</td>
<td>7</td>
<td>0.39</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>xx</td>
<td>size</td>
<td>[0.14, 1]</td>
<td>[−0.02, 0.94]</td>
<td>0</td>
<td>0.22</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>time</td>
<td>[−0.03, 0.97]</td>
<td>[−0.98, −0.15]</td>
<td>6</td>
<td>0.37</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

#### 6.1. Synthesis and interpretation of results

To support the discussions, we rely on a table that summarizes the major results of RQ1, RQ2 and RQ3 by providing different indicators per software system and per performance property. Specifically, Table 4 reports the standard deviation of Spearman correlations (as in RQ1), the minimal and maximal effects of the most influential option (as in RQ2), the average relative difference of performance due to inputs (as in RQ3). Out of the results of Table 4, we can make further observations. First, input sensitivity is specific to both a configurable system and a performance property. For instance, the sensitivity of x264 configurations differs depending on whether bitrate or cpu are considered. Second, there are configurable systems for which inputs threaten the generalization of configuration knowledge, but the performance ratios remain affordable (e.g., SQLITE for q1). Intuitively, one needs a way to assess the level of input sensitivity per system and per performance property. We propose a metric that aggregates both indicators of RQ1, RQ2 and RQ3. We define the score of Input Sensitivity as follows:

$$IS = \frac{1}{4} \cdot |C_{max} - C_{min}| + \frac{1}{25} \cdot \min(Q_2 - 1, \alpha)$$

where $C_{min}$ and $C_{max}$ are the minimal and maximal Spearman correlations $Q_2$ is the median of the performance ratio distribution, and $\alpha$ a threshold representing the maximal proportion of variability due to inputs we can tolerate. The first part of the formula quantifies (in [0, 0.5]), as $|C_{max} - C_{min}|$ is in [0, 2] how the input sensitivity changes the configuration knowledge ($RQ1$ and $RQ2$). For instance, a textbook case of software system with no input sensitivity would have only performance correlations of 1, leading to a first part equal to $\frac{1}{4} \cdot |1 - 1| = 0$. But if the correlations are completely opposite between different inputs, this first part would be equal to $\frac{1}{4} \cdot |1 - (-1)| = \frac{2}{4} = 0.5$. The second part quantifies (in [0, 0.5]) the impact of input sensitivity ($RQ3$) in the actual performance. For instance, a software system with no impact of input sensitivity would have only performance ratios equals to 1, leading to $Q_2 = 1$ and $\frac{1}{25} \cdot \min(Q_2 - 1, \alpha) = 0$. Conversely, for high performance ratios, $Q_2 - 1 > \alpha$, $\min(Q_2 - 1, \alpha) = \alpha$ and $\frac{1}{25} \cdot \min(Q_2 - 1, \alpha) = \frac{\alpha}{25} = 0.5$ IS thus varies between 0 (no input sensitivity) and 1 (high input sensitivity). We compute IS for each couple of systems and performance properties of our dataset, with $\alpha$ fixed at 25% (see Table 4). Empirical evidences show that IS values are robust and trustworthy when using the measurements of 15 inputs or more.17

IS scores are reported in Table 4 as follows: systems and performance properties with scores higher than 0.5 as input-sensitive (lightgray), and those with IS greater than 0.8 as highly input-sensitive (gray). IS scores highlight the input sensitive cases e.g., 0.98 for the time of poppler, 0.84 for the bitrate and the size of x264. Systems like xx or imagemagick exhibit low IS scores that reflect their low sensitivity to inputs. As a small validation, we also

---

compute the IS of \textit{x264} for input videos used in [5]. We retrieve scores of 0.31 and 0.66 for the \textit{time} and the \textit{size} of \textit{x264}.

\subsection{Implications, insights, open challenges}

Our study has several implications for different tasks related to the performance of software system configurations. For each task, we systematically discuss the key insights and open problems brought by our results and not addressed in the state of the art.

\textbf{Tuning configurable systems.} Numerous works aim to find optimal configurations of a configurable system. 

\textbf{Key insights.} Our empirical results show that the best configuration can be differently ranked (see \textit{RQ}_1) depending on an input. The tuning cannot be reused as such, but should be redone or adapted whenever a system processes a new input. Another key result is that it is worth taking input into account when tuning: relatively high performance gains can be obtained (see \textit{RQ}_3).

\textbf{Open challenges.} The main challenge is to deliver algorithms and practical tools capable of tuning the performance of a system, whatever the input. A related issue is to minimize the cost of tuning. For instance, tuning from scratch – each time a new input is fed to a software system – seems impractical since too costly. Approaches that reuse configuration knowledge through e.g., prioritized sampling or transfer learning can be helpful here, but should be carefully assessed w.r.t. costs and actual performance improvements.

\textbf{Performance prediction of configurable systems.} Numerous works aim to predict the performance of an arbitrary configuration.

\textbf{Key insights.} Looking at indicators of \textit{RQ}_1 and \textit{RQ}_2, inputs can threaten the generalization of configuration knowledge. That is, a performance prediction model trained out of one input can be highly inaccurate for many other inputs.

\textbf{Open challenges.} The ability to transfer configuration knowledge across inputs is a critical issue. Transfer learning techniques have been explored, but mostly for hardware or version changes [98, 59] and not for inputs’ changes. Such techniques require measuring several configurations each time an input is targeted. It also requires training performance models that can be reused. Owing to the huge space of possible inputs, this computational cost can be a barrier if systematically applied. A possible direction for reducing measurements’ cost is to group together similar inputs.

\textbf{Understanding of configurable systems.} Understanding the effects of options and their interactions is hard for developers and users yet crucial for maintaining, debugging or configuring a software system. Some works (e.g., [102]) have proposed to build performance models that are interpretable and capable of communicating the influence of individual options on performance.

\textbf{Key insights.} Our empirical results show that performance models, options and their interactions are sensitive to inputs (see indicators of \textit{RQ}_2). To concretely illustrate this, we present a minimal example using SPLConqueror [85] a tool to synthesize interpretable models. We trained two performance models predicting the encoding sizes of two different input videos fed to \textit{x264}. Unfortunately, the two related models do not share any common (interaction of) option.\footnote{See the performance models for the first and the second input videos.} Let us be clear: the fault lies not with SPLConqueror, but with the fact that a model simply does not generalize to any input.

\textbf{Open challenges.} Hence, a first open issue is to communicate when and how options interact with input data. The properties of the input can be exploited, but they must be understandable to developers and users. Another challenge is to identify a minimal set of representative inputs (see \textit{RQ}_4) in such a way interpretable performance models can be learnt out of observations of configurable systems.

\textbf{Effectiveness of sampling and learning strategies.} Measuring a few configurations (a sample) to learn and predict the performance of any configurations has been subject to intensive research. The problem is to sample a small and representative set of configurations and inputs that leads to a good accuracy.

\textbf{Key insights.} A key observation of \textit{RQ}_2 is that the importance of options can vary across inputs. Therefore, sampling strategies that prioritize or neglect some options may miss important observations if the specifics of inputs are not considered. We thus warn researchers that the effectiveness of sampling strategies for a given configurable system can be biased by the inputs and the performance property used.

\textbf{Open challenges.} Pereira \textit{et al.} [5] showed that some sampling strategies are more or less effective depending on the 19 videos and 2 performance properties of \textit{x264}. Kalte-\textit{necker} \textit{et al.} [44] empirically showed that there is no one-size-fits-all solution when choosing a sampling strategy together with a learning technique. We suspect that input sensitivity further exacerbates the phenomenon. Using our dataset, we are seeing two opportunities for researchers: (1) assessing state-of-the-art sampling strategies; (2) designing input-aware sampling strategies \textit{i.e.}, cost-effective for any input.

\textbf{Testing and benchmarking configurable systems.} With limited budget, developers continuously test the performance of configurable systems for ensuring non-regression.

\textbf{Key insights.} Testing software configurations on a single, fixed input can hide several interesting insights related to software properties \textit{(e.g., performance bugs)}. From this perspective, indicators of \textit{RQ}_2 about influences of options should be analyzed and controlled. Similarly, performance
drop (see $RQ_1$) should be handled.

Open challenges. To reduce the cost of measurements, the ideal would be to select a set of input data, both representative of the usage of the system and cheap to measure. We believe our work (see $RQ_4$) can be helpful here. On the $x264$ case study, for the bitrate, we isolate four encoding groups of input videos - see Table 2 in $RQ_2$. Within a group, the videos share common properties, and $x264$ processes them in the same way i.e., same performance distributions ($RQ_1$), same options’ effects ($RQ_2$) and a negligible impact of input sensitivity ($RQ_3$). Automating this grouping could drastically reduce the cost of testing. An approach applicable to any kind of input and configurable software is yet to be defined and assessed.

Detecting input sensitivity. Practitioners and scientists should have the means to determine whether a software under study is input-sensitive w.r.t. the performance property of interest.

Key insights. We propose several indicators (as part of $RQ_1$, $RQ_2$, and $RQ_3$) as well as IS a simple, aggregated score to quantify the level of input sensitivity. Such metrics can be leveraged to take informed decisions as part of the tasks previously discussed.

Open challenges. Detecting input sensitivity has a computational cost. Selecting the right subset of configurations and input data is thus a key issue. Our empirical experiments (see Section 6.1) suggest that a limited percentage of inputs (around 10%) can be used to quantify sensitivity. Other indicators and metrics can also be proposed to quantify sensitivity to inputs. Our study is the first to provide evidence of input sensitivity. We also share data with 1 976 025 measurements that can be analyzed and reused to consolidate configuration knowledge. However, further empirical knowledge is more than welcome to understand the significance of input sensitivity on other software systems and performance properties.

7. Threats to Validity

This section discusses the threats to validity related to our protocol.

Construct validity. Due to resource constraints, we did not include all the options of the configurable systems in the experimental protocol. We may have forgotten configuration options that matter when predicting the performance of our configurable systems. However, we consider features that impact the performance properties according to the documentation, which is sufficient to show the existence of the input sensitivity issue. The use of random sampling also represents a threat, in the sense that the measured configurations could not be representative of a real-world usage of the software systems. To mitigate this threat, we take care of selecting documented and informed options, typically part of custom configurations and profiles, that are supposed to have an effect of performance. We mainly relied on documentation and guides associated to the projects. The validity of the conclusions can depend on the choice of systems under test. In the context of [55], we conducted an additional experiment to ensure the robustness of our results for $x265$, an alternative software to $x264$. Results show that the performance distributions are different from $x264$ to $x265$ (except for size) but the input sensitivity problem holds for $x265$ when it is observed for $x264$.

Internal Validity. First, our results can be subject to measurement bias. We alleviated this threat by making sure only our experiment was running on the server we used to measure the performance of software systems. It has several benefits: we can guarantee we use similar hardware (both in terms of CPU and disk) for all measurements; we can control the workload of each machine (basically we force the machine to be used only by us); we can avoid networking and I/O issues by placing inputs on local folders. But it could also represent a threat: our experiments may depend on the hardware and operating system. To mitigate this, we conducted an additional experiment on $x264$ over a subset of inputs to show the robustness of results whatever the hardware platforms. The measurement process is launched via Docker containers. If this aims at making this work reproducible, this can also alter the results of our experiment. Because of the amount of resources needed to compute all the measures, we did not repeat the process of Figure 2 several times per system. We consider that the large number of inputs under test overcomes this threat. Moreover, related work (e.g., [5] for $x264$) has shown that inputs lead to stable performance measurements across different launches of the same configuration. Finally, the measurement process can also suffer from a lack of inputs. To limit this problem, we took relevant dataset of inputs produced and widely used in their field. For $RQ_4$, we consider oracles when predicting the best configurations for both scenarios, thus neglecting the imprecision of performance models: these results might change on a real-world case. In Section 5, our results are subject to the selection of research papers: since we use and reproduce [72], we face the same threats to validity.

External Validity. A threat to external validity is related to the used case studies and the discussion of the results. Because we rely on specific systems and interesting performance properties, the results may be subject to these systems and properties. To reduce this bias, we selected multiple configurable systems, used for different purposes in different domains.

8. Related Work

In this section, we discuss other related work (see also Section 5).

\textsuperscript{20}See at https://github.com/11esoil/input_sensitivity/blob/master/results/others/x264_x265/x264_x265.ipynb

\textsuperscript{21}See the companion repository at https://github.com/11esoil/input_sensitivity/blob/master/results/others/x264_hardware/x264_hardware.ipynb
Workload Performance Analysis. On the one hand, some work has been addressing the performance analysis of software systems [76, 12, 23, 27, 53, 86] depending on different input data (also called workloads or benchmarks), but all of them only considered a rather limited set of configurations. On the other hand, as already discussed in Section 5, works and studies on configurable systems usually neglect input data (e.g., using a unique video for measuring the configurations of a video encoder). In this paper, we combined both dimensions by performing an in-depth, controlled study of several configurable systems to make it vary in the large, both in terms of configurations and inputs. In contrast to research papers considering multiple factors of the executing environment in the wild [42, 98], we concentrated on inputs and software configurations only, which allowed us to draw reliable conclusions regarding the specific impact of inputs on software variability.

Performance Prediction. Research work has shown that machine learning could predict the performance of configurations [30, 80, 106, 96]. These works measure the performance of a configuration sample under specific settings to then build a model capable of predicting the performance of any other configuration, i.e., a performance model. Numerous works have proposed to model performance of software configurations, with several use-cases in mind for developers and users of software systems: the maintenance and understanding of configuration options and their interactions [83], the selection of an optimal configuration [70, 25, 67], the automated specialization of configurable systems [93, 92]. Input sensitivity complicates their task; since inputs affect software performance, it is yet a challenge to train reusable performance prediction models i.e., that we could apply on multiple inputs.

Input-aware tuning. The input sensitivity issue has been partly considered in some specific domains (SAT solvers [105, 22], compilation [74, 16], video encoding [61], data compression [47], etc.). It is unclear whether these ad hoc solutions are cost-effective. As future work, we plan to systematically assess domain-specific techniques as well as generic, domain-agnostic approach (e.g., transfer learning) using our dataset. Furthermore, the existence of a general solution applicable to all domains and software configurations is an open question. For example, is it always possible and effective to extract input properties for all kinds of inputs?

Input Data and other Variability Factors. Most of the studies support learning models restrictive to specific static settings (e.g., inputs, hardware, and version) such that a new prediction model has to be learned from scratch once the environment change [72]. Janshidi et al. [41] conducted an empirical study on four configurable systems (including SQLite and x264), varying software configurations and environmental conditions, such as hardware, input, and software versions. But without isolating the individual effect of input data on software configurations, it is challenging to understand the existing interplay between the inputs and any other variability factor [54], e.g., the hardware.

9. Conclusion

We conducted a large study over the inputs fed to 8 configurable systems that shows the significance of the input sensitivity problem on performance properties. We deliver one main message: inputs interact with configuration options in non-monotonous ways, thus making it difficult to (automatically) configure a system.

There are also some opportunities when tackling the input sensitivity problem. We have shown it is possible to select a representative set of inputs and thus to greatly reduce the cost of benchmarking software e.g., 1397 → 8 inputs for x264 despite high sensitivity.

Our analysis of the literature showed that input sensitivity has been either overlooked or partially addressed. We have pointed out several open problems to consider related to tuning, prediction, understanding, and testing of configurable systems. In light of the results of our study, we encourage researchers to confront existing methods and explore future ideas with our dataset.

As future work, it is an open challenge to solve the issue of input sensitivity when predicting, tuning, understanding, or testing configurable systems. In particular, a direct follow-up work aim at adapting the current practice of performance models to overcome input sensitivity and train models robusts to the change of input data.

References

[1] Jan De Cock, Aditya Mavlankar, Anush Moorthy, and Anne Aaron: A Large-Scale Comparison of x264, x265, and libvpx – a Sneak Peek. netflix-study (2016)


