Enhancing Thermodynamic Data Quality for Refrigerant Mixtures: Domain-Informed Anomaly Detection and Removal

Laughman, Christopher R.; Deshpande, Vedang M.; Chakrabarty, Ankush; Qiao, Hongtao

TR2024-099 July 16, 2024

Abstract

Next-generation vapor compression cycles will rely upon multicomponent refrigerant mixtures to reduce the climate impact of the working fluids, but the computation of thermodynamic property data for these mixtures is numerically challenging and often results in non-physical anomalies that are present in the output of standard calculation tools. In this paper, we explore two alternative techniques for mitigating the effect of these anomalous points in a reference dataset. The first of these approaches is based upon heteroscedastic Gaussian processes, and builds a statistical model of the property to identify outliers in the reference data. The second uses an estimation method based upon constrained optimization to first detect these outliers and then compute optimal perturbations to the reference data so that the resulting target dataset satisfies domain-informed constraints on the reference data. We demonstrate the efficacy of these methods in computing a target dataset for the refrigerant R454C that is free of anomalies, and which can then be used to build computationally efficient models for use in dynamic cycle simulations.

nternational Refrigeration and Air Conditioning Conference at Purdue 2024

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Enhancing Thermodynamic Data Quality for Refrigerant Mixtures: Domain-Informed Anomaly Detection and Removal

Christopher R. Laughman, Vedang Deshpande, Ankush Chakrabarty, Hongtao Qiao

Mitsubishi Electric Research Laboratories,
Cambridge, MA, USA
{laughman,deshpande,chakrabarty,qiao}@merl.com

ABSTRACT

Next-generation vapor compression cycles will rely upon multicomponent refrigerant mixtures to reduce the climate impact of the working fluids, but the computation of thermodynamic property data for these mixtures is numerically challenging and often results in non-physical anomalies that are present in the output of standard calculation tools. In this paper, we explore two alternative techniques for mitigating the effect of these anomalous points in a reference dataset. The first of these approaches is based upon heteroscedastic Gaussian processes, and builds a statistical model of the property to identify outliers in the reference data. The second uses an estimation method based upon constrained optimization to first detect these outliers and then compute optimal perturbations to the reference data so that the resulting target dataset satisfies domain-informed constraints on the reference data. We demonstrate the efficacy of these methods in computing a target dataset for the refrigerant R454C that is free of anomalies, and which can then be used to build computationally efficient models for use in dynamic cycle simulations.

1. INTRODUCTION

Thermodynamic property models of refrigerants play an essential role in dynamic physics-based models of vapor-compression cycles used for model-based design to reduce development times and increase system performance. As these refrigerant models enforce algebraic constraints that describe the nonlinear relations between the property variables, such as pressure, temperature, density, and specific enthalpy, the accuracy of the overall system model is strongly dependent upon the accuracy of the property models. Model simulation speed is also governed by the speed of the property models in many cases; for example, refrigerant density functions were called more than 400 million times over a 130 second period in one of our dynamic simulations of an on/off sequence for a simple vapor-compression cycle over a 2 hour window.

Recent awareness about the environmental impact of refrigerants have brought issues of refrigerant selection to the forefront of system design considerations (McLinden & Huber, 2020). While previous generations of refrigerants have been effective in space conditioning applications, they often contribute significantly to climate change. These trends have motivated significant research and study into multicomponent refrigerant mixtures which have much lower global warming potential than conventional fluids.

Unfortunately, these refrigerant mixtures tend to be harder to describe using first principles methods than pure refrigerants, due to the numerical challenges often encountered during model implementation. The overall structure of the methods used in reference software such as REFPROP (Lemmon et al., 2018) generally involves a series of nested iterative root-finding computations that enforce physical constraints that apply in the thermodynamic phase space, e.g., the equality of pressures, temperatures, and fugacities in flash calculations (Span, 2000). As is the case for any nonlinear root-finding problem, the success of these algorithms is highly dependent upon the shape of the function under study, and these algorithms often rely on carefully tuned parameters to avoid local minima or limit cycles during convergence. Despite the care taken in the implementation of these numerical methods, however, these property calculation routines thus sometimes return erroneous values at specific state points over the range of operation due to poor local numerical behavior.

We can see an example of this behavior in Figure 1, which illustrates the density surface ρ computed by REFPROP over a wide domain of conditions for the refrigerant mixture R454C as a function of pressure P and specific enthalpy h. This figure illustrates a heatmap in which the color is proportional to the refrigerant density, and where the bubble

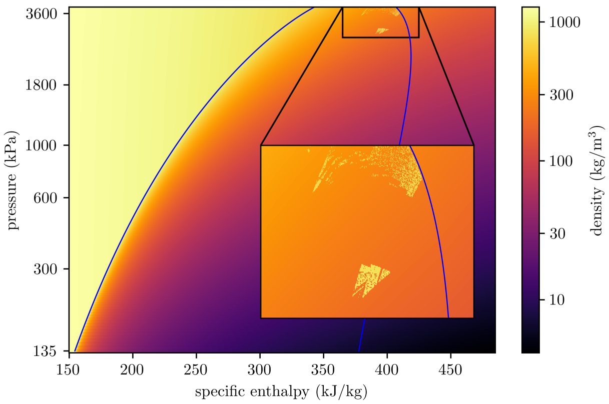

Figure 1: Reference density surface for R454C obtained from REFPROP; inset shows detail of anomalous datapoints close to critical point.

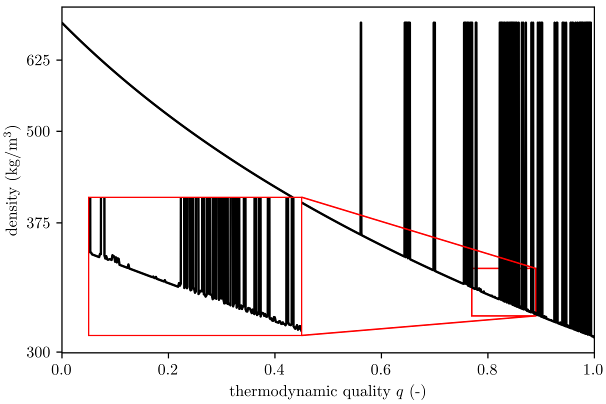

Figure 2: Density $\rho(q)$ in the two-phase region at a pressure of 3.785 MPa; inset shows detail of saturation line with both small and large anomalies.

and dew lines are drawn in bright blue. While much of the density surface is smooth, anomalous data is visible in the two-phase region at pressures close to the critical point. A detailed view of this anomalous data is provided in the inset plot, where it is clear that the anomalies are not uniformly distributed in the $P-h$ space, but are instead scattered throughout this region.

An additional view of this output can be seen in Figure 2, which illustrates $\rho(q)$ at a pressure of 3.785 MPa, where q is the thermodynamic quality in the two-phase region, defined as

$$q = \frac{h - h_{bub}(P)}{h_{dew}(P) - h_{bub}(P)}. \tag{1}$$

The anomalous density outputs that result from nonconvergence of the iterative algorithms can be clearly seen in the figure, and are unevenly distributed throughout the two-phase region at values of q greater than 0.5. REFPROP provides an indication of some these errors by raising an flag when specific conditions are met (e.g., hitting an iteration limit), but many of these anomalous points are not accompanied by an error flag. The inset figure illustrates an additional aspect of these outputs; in addition to the large-scale anomalies, small-scale non-physical variations are also present in the output. Such variations will result in large spikes in the density derivatives, which are often used in the formulation of the mass and energy conservation equations, that cause dynamic simulations to yield inaccurate predictions or simply fail.

We note that the requirements for a general thermodynamic property calculation package, such as REFPROP, differ from the requirements for property models used in dynamic system-level design. REFPROP must accommodate arbitrary mixtures of fluids and produce physically reasonable estimates of the fluid properties over wide ranges of operating conditions; this motivates the use of first-principles models due to a paucity of reference data for hypothetical mixtures under consideration. On the other hand, dynamic system-level design is usually conducted on a limited set of fluids that have been evaluated extensively and have been shown to meet thermodynamic criteria that are amenable for practical use. While REFPROP was designed to address the fluid selection problem, the requirements of the system design problem are quite different and motivate the development of specialized thermodynamic fluid property models for this application.

For the purposes of model-based system design, we seek to develop computationally efficient refrigerant property models that satisfy the demanding requirements of dynamic cycle simulations. We therefore seek a means for operating on *reference* data as generated by a tool such as REFPROP, which may contain anomalous data points, by removing those anomalies and replacing them with estimated values that better match the expected physical properties of the fluid. This operation thus results in a set of *target* data that can be used to build non-iterative models using interpolatory methods that meet stringent requirements of speed, accuracy, consistency, and smoothness. This last property of smoothness is particularly important for the fluid model, as it ensures that the derivatives are only discontinuous at points where this is physically expected, e.g., saturation lines, and that there are no discontinuities in the derivatives elsewhere on the property surfaces. This is strongly dependent upon obtaining a sufficiently smooth set of target data from which the model is built.

Different modeling approaches for these applications have been studied in a substantial body of prior work, but no extant papers describe systematic methods to accommodate anomalous reference data in developing thermodynamic property models. Much of the relevant literature examines the use of different approximation methods to represent the property curves and surfaces to gain computational advantages; for example, Miyagawa & Hill (2001) develop table-based Taylor series expansions, while Kunick (2018) develops similar spline-based approaches to represent the property surfaces in a computationally efficient manner. Li et al. (2018) extend these spline-based methods and applies them specifically to vapor-compression cycles, while Aute & Radermacher (2014) use Chebyshev polynomials to approximate these surfaces for similar applications.

In this work, we describe two methods for identifying and eliminating anomalous reference data for later use in constructing thermodynamic property models. This paper does not focus on the details of the property models for use in dynamic simulation, but rather upon methods for manipulating the source data in a physically consistent manner to eliminate the effect of numerical errors that prevent the construction of computationally efficient model formulations. This source data could then be used to construct a variety of different property models, such as those described by Aute & Radermacher (2014), Li et al. (2018), or Laughman & Qiao (2021).

This paper proceeds in Section 2 by developing a data-driven approach using heteroscedastic Gaussian processes (GPs) to characterize the uncertainty in the reference data and label points that are outside a 3σ band as anomalies, after which the remaining data is used to construct the target data. This method is able to successfully identify the large-scale anomalies present in the data, but is unable to eliminate the effect of the small-scale anomalies that cause non-physical oscillations in the density derivatives. We address this shortcoming in Section 3, in which we develop an constrained optimization-based approach that computes optimal perturbations to the data to satisfy thermodynamically motivated requirements. We then demonstrate that this approach can successfully eliminate all extant anomalies in the data, and produces target data from which fast dynamic property models can be built. Finally, Section 4 summarizes the contributions of this work and identifies directions for future research.

2. HETEROSCEDASTIC GAUSSIAN PROCESSES

For the purposes of the following work, we assume that we have access to thermodynamic data samples from REFPROP for the fluid R454C, consisting of refrigerant density values at corresponding values of thermodynamic quality. We restrict our study to that of density, as the conservation equations used to describe dynamic model behavior are often written in a form that includes the density derivatives, but the described methods are general and can be extended to other thermodynamic properties, such as temperature or specific entropy. On an empirical basis, we have observed that the numerical computation of density at pressures above approximately 70% of the critical pressure often yields anomalous density values for low-GWP refrigerant mixtures. In this section, we describe a probabilistic machine learning framework for automatically identifying and removing the anomalous density values, and subsequently replacing those values with more realistic estimates of density using function approximators.

We thus propose the use of Gaussian processes (GPs), which are a family of machine learning methods used for probabilistic regression (Williams & Rasmussen, 2006), to eliminate the anomalies from the reference data. Let $\mathcal{Q}$ denote the set of thermodynamic qualities for which the density values are to be calculated. We also assume that data pairs are available from REFPROP, where each pair consists of a quality q_i and a corresponding density ρ_i, and $i = 1, \ldots, n$, where $n \in \mathbb{N}$ is the size of the dataset available for regression. For a given quality $q \in \mathcal{Q}$, the predicted density ρ is estimated using the equation $\rho = \mathcal{GP}(q) + \varepsilon$, where $\mathcal{GP}(q)$ is a function generated from a Gaussian process prior equipped with a mean function $\bar{\mu}(q)$ and a covariance function $k(q, q')$. The covariance or kernel function k is positive definite, with parameterized families such as the Gaussian or Matérn being typical choices (Binois et al., 2018).

In classical (homoscedastic) GP regression, the term ε denotes an i.i.d. Gaussian noise term with zero mean and variance σ_ε^2. However, in this work, we can leverage domain knowledge to construct heteroscedastic GPs (Kersting et al., 2007; Goldberg et al., 1997) that are amenable to automatically learn which data samples in $\{\rho_k\}_{k=1}^n$ are anomalous.

The heteroscedastic GP (hGP) assumes that the noise distribution depends on the thermodynamic quality, that is, $\rho = \mathcal{GP}(q) + \varepsilon(q)$, where the noise distribution is a zero-mean Gaussian with quality-dependent covariance $\sigma_\varepsilon^2(q)$. Consequently, one can write the mean and variance

$$\begin{aligned} \mu(q) &= \bar{\mu}(q) + k_n(q)(K_n + \Sigma_n)^{-1}\rho_n, \\ \sigma^2(q) &= k(q,q) + \sigma_\varepsilon^2(q) - k_n(q)^\top (K_n + \Sigma_n)^{-1} k_n(q), \end{aligned} \tag{2}$$

of the posterior predictive distribution for a desired quality value $q \in \mathcal{Q}$. Here, $\rho_n := \begin{bmatrix} \rho_1 & \rho_2 & \cdots & \rho_n \end{bmatrix}^\top$ is a column vector of the available density data values, and the kernel matrices are given by

$$k_n := \begin{bmatrix} k(q,q_1) \\ k(q,q_2) \\ \vdots \\ k(q,q_n) \end{bmatrix}, \; K_n = \begin{bmatrix} k(q_1,q_1) & k(q_1,q_2) & \dots & k(q_1,q_n) \\ k(q_2,q_1) & k(q_2,q_2) & \dots & k(q_2,q_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(q_n,q_1) & k(q_n,q_2) & \dots & k(q_n,q_n) \end{bmatrix}, \text{ and } \Sigma_n = \begin{bmatrix} \sigma_\varepsilon^2(q_1) & 0 & \dots & 0 \\ 0 & \sigma_\varepsilon^2(q_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_\varepsilon^2(q_n) \end{bmatrix}.$$

Training the hGP involves computing suitable parameters that define the kernel function k by maximizing a log-likelihood loss function; more details about the training procedure are provided in Section 2 of Binois et al. (2018).

On the basis of our empirical experience and the fact that numerous nested Newton-Raphson iterations are usually required to compute the thermodynamic equilibria for mixtures in boiling or condensing conditions (Span, 2000), we expect that the anomalous data will occur in a neighborhood of the quality range $[0, 1]$. We embed this information in the hGP training process, thereby incorporating a domain-informed learning aspect to the machine learning module, by assigning different prior noise covariance values to different ranges of quality. In particular, we construct the matrix Σ_n above with small $\breve{\sigma}_\varepsilon(q)$ values for qualities outside a user-defined $\Delta := [-\delta, 1 + \delta]$, and large $\hat{\sigma}_\varepsilon$ within the range Δ. By assigning a larger variance within Δ, we prevent the hGP overfitting the data in that range by generating a mean function that does not have to graze the anomalous samples. The anomalous samples can instead be explained by the hGP inducing a wide uncertainty band around the mean function. Conversely, we choose the noise covariance to be small in the quality range $\mathcal{Q} \setminus \Delta$ because we want the mean function to closely resemble the data, with a small uncertainty band around it, as we do not expect the data to contain anomalies. Once the hGP is trained, one can use the uncertainty quantified by the posterior covariance σ^2 to identify anomalous samples.

Algorithm 1 Data Cleaning Algorithm

```
Require: ${q_i, \rho_i}_{i=1}^N$ ← Training dataset of qualities and densities for refrigerant at a fixed pressure
Require: $\Delta$ ← Expected range of qualities where anomalous samples reside
Require: $k$ ← type of kernel for hGP
Require: $\breve{\sigma}_\varepsilon, \hat{\sigma}_\varepsilon$ ← variances at quality locations for hGP
Require: $n_r, n_s$ ← number of randomized runs and size of random subset
Select fixed-noise Gaussian likelihood parameterized by $\breve{\sigma}_\varepsilon$ for $q \notin \Delta$ and $\hat{\sigma}_\varepsilon$ for $q \in \Delta$
Train hGP with data and heteroscedastic kernel with fixed-noise likelihood using Adam optimizer
for $i = 1 : n_r$ do
    Select random subset of data of size $n_s$
    $\mu(q), \sigma(q)$ ← predicted posterior distribution using hGP
    Compute 99% confidence interval (CI)

    for $j = 1 : n_s$ do
        Flag $j$-th datapoint as anomalous if outside 99% CI
    end for
end for
Identify indices of data that are consistently flagged as anomalous
$\mathcal{I}$ ← index of data points after anomalous samples removed
    return Cleaned dataset ${q_i, \rho_i}_{i \in \mathcal{I}}$
```

The data cleaning algorithm is constructed as follows. We first generate n_r random subsets of the training data, which may be of different sizes or (for simplicity) a fixed size $n_s < n$. Recall that each data sample in the subset is a pair (q, ρ).

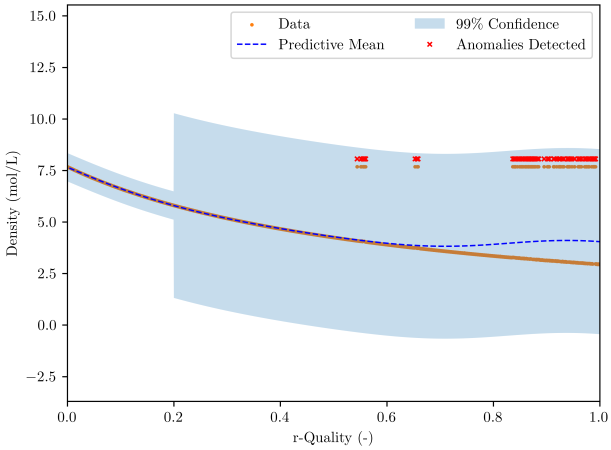

Figure 3: Prediction and anomaly identification with the trained hGP.

For each of these subsets, we use the trained hGP to predict the mean and variance of corresponding density values, using Equation 2. Since the posterior prediction is Gaussian, we can easily compute the 99% confidence interval (CI) around the mean density prediction by using $\mu(q) \pm 3\sigma(q)$. If any density data value in the random subset lies outside of the 99% CI, we then label that data sample anomalous. By repeating this procedure for n_r random subsets, we can obtain n_r sets of anomalous density values. We argue that a truly anomalous density value will consistently be labeled anomalous across most random subsets because it will consistently lie outside the confidence regions predicted by the hGP. As random subsets of sufficiently large size n_s are expected to mostly contain non-anomalous data, we expect that the mean predictive function generated by the hGP should not have large jumps and will not be able to fit the anomalous density values well. Since the hGP allows some more noise within Δ, it accepts some small noise around the density values (quantified by the 99% CI), but if a density value is outside even this region, it cannot be explained by the distribution induced by the hGP, and therefore, is most likely an anomalous data sample.

At the end of the above procedure, one obtains a list of data points that have appeared in the random subsets, as well as a count of how many times each such data point has been labeled anomalous. Based on this count, these points can be removed. If the original dataset is large, and the number of identified anomalous points is small, all points labeled anomalous at least once could then be removed. Otherwise, one could choose an integer threshold η and remove reference data that has been labeled anomalous $> \eta$ times. Whichever heuristic is used, one can obtain an index set $\mathcal{I} \subset \{1, \ldots, N\}$ such that the dataset $\{q_i, \rho_i\}_{i \in \mathcal{I}}$ contains only data samples that have not been consistently labeled anomalous using our hGP-based anomaly detection procedure.

Figure 3 illustrates the ability of this method to identify anomalies in the reference data for density of R454C. In this figure, the the dotted orange lines indicate the data obtained from REFPROP, while the blue dashed line represents the hGP predictive mean and the uncertainty band represents the 99% confidence interval. The hGP predictive mean fits the data well by throwing out the identified anomalies shown with the red ×.

While this method is clearly effective at identifying the large-scale anomalies present in the output of the REFPROP density calculations, it is not as effective at identifying the small-scale anomalies. This is due to the fact that the small-scale anomalies are close in value to the non-anomalous data, so it is difficult for the statistical methods to sufficiently differentiate between these classes of data points. In addition, this approach does not readily provide a mechanism for replacing the anomalous data with improved estimates. While the predictive mean from the GPs could be used, it is clear from Figure 3 that there is a significant discrepancy between the predictive mean and the non-anomalous data. We therefore consider an alternative method for detecting anomalies in the reference data and generating estimates for use in the target data set, as described in the following section.

3. ESTIMATION BY CONSTRAINED OPTIMIZATION

Given the limitations of the GP-based method for identifying small-scale anomalies in the reference data, we propose a second optimization-based approach for estimating anomaly-free source data that consists of two steps. First, we use statistical methods to identify both large-scale and small-scale anomalous points from the reference data. Once the set of anomalous data points have been identified, we subsequently replace these points with values that are determined by solving an optimization problem subject to physics-informed constraints on thermodynamic properties. As a result, the generated target data exhibits proper thermodynamic behavior while remaining close to the original reference data.

This method begins by identifying the anomalous reference data from the full set $\{q_i, \rho_i\}_{i=1}^{N}$ available from REFPROP for a thermodynamic property ρ as a function of thermodynamic quality q, i.e., $\rho_i = \rho(q_i)$, due to the fact that q is normalized between 0 and 1, though ρ_i can also be calculated directly from pressure P_i and specific enthalpy h_i. We assume that the quality values are uniformly spaced within the interval $[q_1, q_N]$. As before, ρ represents refrigerant density, though this method could be used for other quantities as well. This approach for identifying of anomalous data points is outlined in Algorithm 2 and is described below.

Algorithm 2 Identification of anomalous data

```
Require: $\{q_i, \rho_i\}_{i=1}^{N}$ ← Reference data from REFPROP
Require: $L$ ← window length
Require: $\varepsilon$ ← threshold value
$\mathcal{A} \leftarrow \varnothing$
for $k = 1 : K$ do                                          ▷ K=N/L
    $\mathcal{D}_k := \{\rho_i\}_{i=(k-1)L+1}^{kL}$
    $R_k = \texttt{acf}_1\big(\, \texttt{F}_1(\mathcal{D}_k)\, \big)$
    if $R_k < \varepsilon$ then
        $\mathcal{A} \leftarrow \mathcal{A} \cup \big\{(k-1)L+1, \cdots kL\big\}$
    end if
end for
return Set of anomalous data indices $\mathcal{A}$
```

We consider K non-overlapping windows or segments containing L consecutive data points of the series $\mathcal{D} := \{\rho_i\}_{i=1}^{N}$ such that $N = KL$. The first window contains the data points $\{\rho_i\}_{i=1}^{L}$, the second window contains the data points $\{\rho_i\}_{i=L+1}^{2L}$, and so on. The number of data points L in a window is referred to as *window length*, and typically $L << N$. For every window, we then determine if there are any anomalous data points within that window based on a statistical metric. If a window is found to have anomalous data, all data indices within that window are flagged as anomalous.

We adapt the lag-1 auto-correlation function ($\texttt{acf}_1$) of the first difference of data values within a window as the statistical metric for identifying anomalous data. For the k^{th} window $\mathcal{D}_k := \{\rho_i\}_{i=(k-1)L+1}^{kL}$, we calculate

$$R_k = \texttt{acf}_1\big(\, \texttt{F}_1(\mathcal{D}_k)\, \big) \tag{3}$$

where $\texttt{F}_d(\cdot)$ denotes the d^{th} forward difference of the data series. The $\texttt{acf}_1$ of an arbitrary data series of real numbers $\mathcal{Z} := \{z_i\}_{i=M}^{N}$ is defined as follows

$$\texttt{acf}_1(\mathcal{Z}) := \frac{\sum_{i=M+1}^{N}(z_i - \bar{z})(z_{i-1} - \bar{z})}{\sum_{i=M}^{N}(z_i - \bar{z})^2}$$

where $\bar{z}$ denotes the mean of the data series.

By definition, R_k lies within the interval $[-1, 1]$. Values close to 1 indicate that the series is smoothly varying; while values close to -1 indicate that the series is jagged in the following sense: if a point is above the mean, the next point is likely to be below the mean by approximately the same amount. As a result, negative values of R_k indicate jagged data series with sharp increments and decrements in the data values. We utilize this property to identify outliers in the

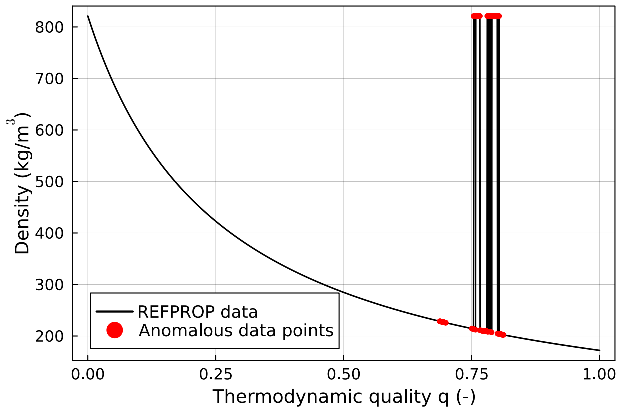

Figure 4: Anomalous reference density data generated by REFPROP at 3.015 MPa. Circular markers show the anomalous data points identified using Algorithm 2.

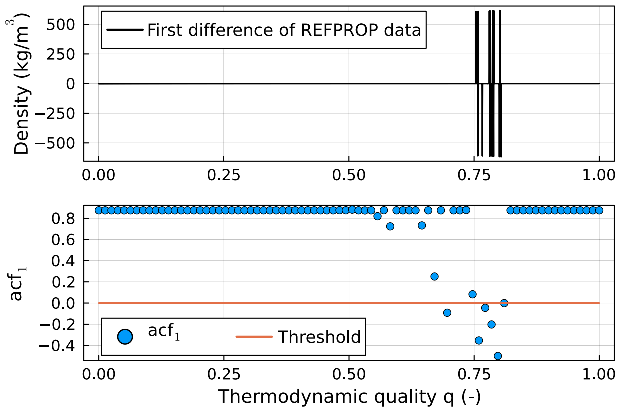

Figure 5: *(Top)* First difference of reference density data at 3.015 MPa. *(Bottom)* $\mathtt{acf}_1$ values calculated for $K = 80$ windows with $L = 25$.

reference data; if there is an outlier in a window, then its first difference will contain values that are above and below the mean by about the same amount, thus yielding a negative value for R_k. In practice, we categorize a window and all data indices within that window as anomalous if $R_k < \varepsilon$ where ε is a prespecified threshold.

This approach to anomaly detection is illustrated in Figures 4 and 5. Figure 4 shows the anomalous reference density data generated from REFPROP. The top plot of Figure 5 shows the first difference of the data, whereas the bottom plot shows $\mathtt{acf}_1$ values for different data windows. Sudden changes in the first difference correspond to negative values of $\mathtt{acf}_1$. For this illustration, the threshold is set to be $\varepsilon = 0$ so that data windows with negative lag-1 autocorrelations are flagged as anomalous, as emphasized in Figure 4 by markers.

Given the anomalous reference data series $\mathcal{D} := \{\rho_i\}_{i=1}^{N}$ and the set of anomalous data indices $\mathcal{A}$ that are identified via Algorithm 2, we eliminate those data points and estimate a cleaned data series $\hat{\mathcal{D}} := \{\hat{\rho}_i\}_{i=1}^{N}$ by solving the optimization problem

$$\begin{aligned} &\min_{\hat{\mathcal{D}}} \sum_{i=1}^{N} \gamma_i(\rho_i - \hat{\rho}_i)^2 + \mathrm{L}_{\mathrm{reg}}(\hat{\mathcal{D}}) \\ &\text{subject to } \hat{\mathcal{D}} \in \mathcal{C} \\ &\gamma_i = 0\ \forall i \in \mathcal{A}, \quad \gamma_i = 1 \text{ otherwise,} \end{aligned} \tag{4}$$

where the weights γ_i corresponding to anomalous data points are set to zero, $\mathrm{L}_{\mathrm{reg}}(\hat{\mathcal{D}})$ denotes a suitable regularization penalty on $\hat{\mathcal{D}}$, and $\mathcal{C}$ denotes a constrained set.

As the first and second derivatives of thermodynamic properties are often used in the formulation of the mass and energy conservation equations, and are thus used by differential equation solvers to advance simulations from an initial condition, the derivatives must exhibit some degree of smoothness and be free of sharp discontinuities except on the saturation curves. We thus induce smoothness on derivatives of up to second order by defining a suitable regularization penalty $\mathrm{L}_{\mathrm{reg}}(\hat{\mathcal{D}})$ as follows

$$\mathrm{L}_{\mathrm{reg}}(\hat{\mathcal{D}}) = \lambda_1 \tilde{\mathrm{R}}(\mathtt{F}_1(\hat{\mathcal{D}})) + \lambda_2 \tilde{\mathrm{R}}(\mathtt{F}_2(\hat{\mathcal{D}})) \tag{5}$$

where $\tilde{\mathrm{R}}(\cdot)$ denotes *roughness* of a data series and the two terms denote roughness of first and second difference of $\hat{\mathcal{D}}$ weighted by the weights λ_1 and λ_2. As the integral of the squared second derivative is conventionally used as a measure of the roughness of a curve, we analogically use an equivalent measure of roughness for a discrete series $\mathcal{Z} := \{z_i\}_{i=1}^{N}$ as given by the sum of its squared second differences, i.e.,

$$\tilde{\mathrm{R}}(\mathcal{Z}) := \sum_{i=1}^{N-2} (z_i'')^2 \tag{6}$$

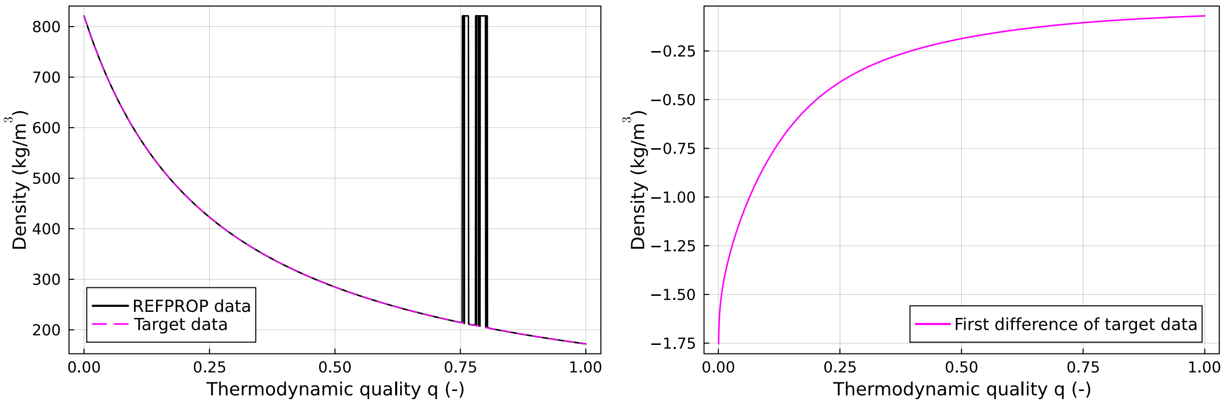

Figure 6: *(Left)* Anomalous reference density data generated by REFPROP at 3.015 MPa and cleaned data estimated by solving (4). *(Right)* First difference of the target data.

where $\mathcal{Z}'' = \{z_i''\}_{i=1}^{N-2} := \mathtt{F}_2(\mathcal{Z})$. Since the roughness is defined in terms of the second difference, we note that the regularization $\mathrm{L}_{\mathrm{reg}}(\hat{\mathcal{D}})$ essentially necessitates the computation of the third and fourth order difference of the optimization variable $\hat{\mathcal{D}}$.

While the regularization improves the smoothness of cleaned data and its derivatives, the final solution may benefit from additional explicit constraints. For example, we know that a cleaned density curve must monotonically decrease with increasing quality. Such a constraint can be incorporated in the optimization problem (4) by specifying a constraint set $\mathcal{C} : \{\hat{\mathcal{D}} \,|\, \mathtt{F}_1(\hat{\mathcal{D}}) < 0\}$. Such constraints can originate from fundamental physical laws and/or they can be inferred from anomaly-free reference data. Such constraints for density include $\mathtt{F}_1(\hat{\mathcal{D}}) < 0$, $\mathtt{F}_2(\hat{\mathcal{D}}) > 0$, $\mathtt{F}_3(\hat{\mathcal{D}}) < 0$, and $\mathtt{F}_4(\hat{\mathcal{D}}) > 0$. By incorporating these constraints, we can ensure that the cleaned data is consistent with the other anomaly-free data generated by REFPROP in terms of monotonicity and the convexity/concavity of property curves.

One advantageous aspect of this data cleaning approach is that the optimization problem is a quadratic program (QP) subject to linear constraints, and as such can be solved using specialized QP solvers as well as some general purpose optimization solvers, including SCS (O'Donoghue et al., 2023), Mosek (ApS, 2024), and Ipopt (Wächter & Biegler, 2006). The stringent requirements of this problem still may pose numerical challenges to solving this problem in practice, including a large problem size, numerical ill-conditioning, and a large number of solver iterations needed for convergence to the optimal solution. These difficulties are further exacerbated by the fact that the regularization $\mathrm{L}_{\mathrm{reg}}(\hat{\mathcal{D}})$ computes higher order differences that can be close to or smaller than the best numerical tolerances supported by many solvers. These numerical challenges also accompany the enforcement of higher derivative constraints. While some of these numerical challenges can be alleviated by normalizing and scaling the reference data, tuning the weights λ_1, λ_2 by trial and error, and adjusting tolerances and maximum iterations of the solver, the nature of such customization will depend strongly upon the method's implementation.

The benefit of applying these methods can be seen by considering the results illustrated in the plots shown in Figure 6. The left plot in Figure 6 illustrates both the reference density and the target density for $q \in [0, 1]$ at 3.015 MPa, where it is clear that the large-scale anomalies present in the reference data from REFPROP have been eliminated in the target data. This is confirmed in the right plot of this same figure, which illustrates the first difference of the target density data; the apparent smoothness of this first difference indicates that there are no abrupt changes in the underlying data.

Figures 7 and 8 further demonstrate the efficacy of this cleaning method on the same reference data as was used to generate Figures 1 and 2. The anomalies clearly present in the reference data over the entire $P - h$ domain have been eliminated, as suggested by the comparison between the inset in both figures. The reference and target densities at 3.785 MPa for $q \in [0, 1]$ illustrate the much smoother behavior of the target density. The values of the target data over the region in the inset are estimated and may slightly differ from the ”non-anomalous” reference data, but these curves manifest physically realistic characteristics (e.g., smoothness of derivatives) that are not exhibited by the reference data and are therefore more useful in a practical context.

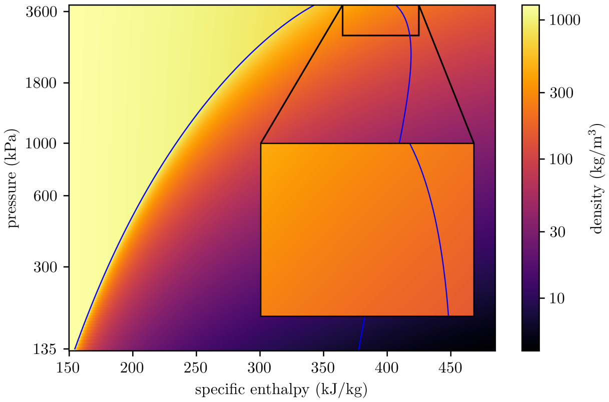

Figure 7: Target density surface for R454C after optimization-based cleaning; inset shows absence of anomalies close to critical point.

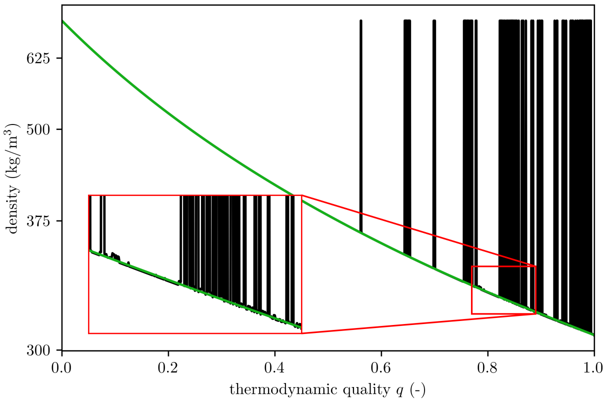

Figure 8: Density $\rho(q)$ in the two-phase region at a pressure of 3.785 MPa; inset shows detail of both reference and target saturation curves.

The fidelity of the target data to the reference data may also be computed for assessment of its accuracy, as all of the target data was potentially adjusted during the constrained optimization phase of the cleaning process. Such an assessment requires some care to avoid comparison to anomalous points in the reference data set. We therefore marked and eliminated all of the anomalous reference data to facilitate a representative comparison between these datasets. Two statistical metrics were computed for the remaining data, including the absolute relative error (ARE) and the RMSE. The mean and maximum values of the ARE over the non-anomalous points in the domain, as computed by

$$\mathrm{ARE}(\rho, \hat{\rho}) = \left| \frac{\rho_i - \hat{\rho}_i}{\rho_i} \right|, \tag{7}$$

was used to identify the characteristic deviations between the reference and target datasets; the mean ARE was 4.432e-6, while the maximum ARE over all data indices was 3.207e-3. Moreover, the RMSE in comparing the nonanomalous reference data to the target data is 0.0224 kg/m^3, suggesting that the target data is quite close to the reference data over the domain.

4. CONCLUSIONS & DISCUSSION

Accurate and computationally suitable models for refrigerant mixtures are important for the model-based design of next-generation vapor compression cycles, but standard reference software for computing thermodynamic properties for these mixtures can produce anomalous data that must be eliminated before it is used to build property models for dynamic simulations. We describe two candidate methods for identifying and eliminating these erroneous data points, and demonstrate the efficacy of these methods on the mixture R454C.

This work serves as an initial exploration of anomaly detection and mitigation methods for thermodynamic property data, but a wide range of topics could be explored in this area in the future. In particular, while these methods should theoretically be applicable to other property variables, such as temperature or specific entropy, efforts to demonstrate the application of these anomaly mitigation methods for those variables would be valuable in its own right. There is also a rich literature related to denoising and anomaly elimination in signal processing and computer vision that is highly relevant to these problems, and further exploration into such methods is likely to provide new insights and approaches for this challenging problem.

REFERENCES

ApS, M. (2024). The MOSEK optimization toolbox for MATLAB manual. version 10.1. [Computer software manual]. Retrieved from `http://docs.mosek.com/latest/toolbox/index.html`

Aute, V., & Radermacher, R. (2014). Standardized polynomials for fast evaluation of refrigerant thermophysical properties. In *International Refrigeration and Air-Conditioning Conference at Purdue.*

Binois, M., Gramacy, R. B., & Ludkovski, M. (2018, 10). Practical heteroscedastic Gaussian process modeling for large simulation experiments. *Journal of Computational and Graphical Statistics*, *27*, 808-821. Retrieved from `https://www.tandfonline.com/doi/full/10.1080/10618600.2018.1458625` doi: 10.1080/10618600.2018.1458625

Goldberg, P. W., Williams, C. K., & Bishop, C. M. (1997). Regression with input-dependent noise: a Gaussian process treatment. In *Proc. 10th int. conf. on neural information processing systems* (pp. 493–499).

Kersting, K., Plagemann, C., Pfaff, P., & Burgard, W. (2007). Most likely heteroscedastic Gaussian process regression. In *Proc. 24th int. conf. on machine learning (icml)* (pp. 393–400).

Kunick, M. (2018). *Fast calculation of thermophysical properties in extensive process simulations with the spline-based table look-up method (SBTL)* (No. 618). Fortschritt-Bericchte VDI.

Laughman, C., & Qiao, H. (2021). Patch-based thermodynamic property models for the subcritical region. In *International Refrigeration and Air-Conditioning Conference at Purdue.*

Lemmon, E. W., Bell, I., Huber, M. L., & McLinden, M. O. (2018). *NIST Standard Reference Database 23: Reference Fluid Thermodynamic and Transport Properties-REFPROP, Version 10.0, National Institute of Standards and Technology.* Retrieved from `https://www.nist.gov/srd/refprop` doi: https://doi.org/10.18434/T4/1502528

Li, L., Gohl, J., Batteh, J., Greiner, C., & Wang, K. (2018). Fast calculation of refrigerant properties in vapor compression cycles using spline-based table look-up method (SBTL). In *Proceedings of the American Modelica Conference 2018.*

McLinden, M., & Huber, M. (2020). (R)Evolution of refrigerants. *Journal of Chemical and Engineering Data*, *65*, 4176-4193. doi: 10.1021/acs.jced.0c00338

Miyagawa, K., & Hill, P. (2001, Jul). Rapid and accurate calculation of water and steam properties using the tabular taylor series expansion method. *Journal of Engineering for Gas Turbines and Power*, *123*, 707-712.

O'Donoghue, B., Chu, E., Parikh, N., & Boyd, S. (2023, November). *SCS: Splitting conic solver, version 3.2.4.* `https://github.com/cvxgrp/scs`.

Span, R. (2000). *Multiparameter equations of state*. Springer-Verlag.

Williams, C. K., & Rasmussen, C. E. (2006). *Gaussian processes for machine learning* (Vol. 2) (No. 3). MIT press Cambridge, MA.

Wächter, A., & Biegler, L. (2006). On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*(106), 25–57. Retrieved from `https://doi.org/10.1007/s10107-004-0559-y`