ACCURACY ASSESSMENT OF A THREE-DIMENSIONAL STRAIN-BASED MODEL FOR DENT ANALYSIS OF PIPES

by

Yueying Li

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Structural Engineering

Department of Civil and Environmental Engineering

University of Alberta

ABSTRACT

Pipelines are the safest and most efficient way to transport oil and gas products throughout the world. Thus, pipelines have to traverse long distances and are typically buried underground which are susceptible to damages with use. Potential threats to the integrity of a pipeline include metal loss, cracking, dents, or the interaction of any of these. Among them, dents, defined as permanent inward plastic deformations localized in the pipe wall, can occur due to external impact, such as strike by construction equipment or settlement over rocks. The dent could further cause coating damage and in turn accelerate growth of corrosion or make the pipe more susceptible to cracking in the deformed area. It becomes necessary to assess the severity of dents in order to prioritize resource allocation in implementing management strategies.

The Canadian pipeline code, CSA Z662-16, specifies that plain dents with depth greater than 6% of the nominal pipe diameter should be excavated and repaired. Because of the geometry, dents associated with localized strain and stress distribution have a greater potential to form and propagate cracks under cyclic pressurization when a pipeline is in operation. This is why they are of greater concern as a dent might fall below the codified deformation limits while violating the localized plastic strain or stress limits. As an alternative to the traditional depth-based criteria, the American Society of Mechanical Engineers Standard for gas pipelines, ASME B31.8-16, presents a set of non-mandatory analytical equations to predict the maximum strain in dents. More recently, numerical modeling via finite element analysis (FEA) has been proposed in literature as an accurate dent-assessment technique.

In-line inspection (ILI) tools can take readings of the inner diameter of the pipe and indicate the location, shape, and size of dents. However, there is no universal dent assessment criterion that can take all dent features into account. The challenge now facing pipeline operators is that a large

number of dents are being reported by ILI tools. Although FEA can model the full geometry of the dents and pipes, it is impossible to process large numbers of dents. Recently, the authors' research group developed a robust but much simplified analytical model to evaluate the strains in dented pipes based solely on data obtained from inline inspection devices. When the strain distribution predicted using the analytical model is benchmarked against the strains by nonlinear FEA they showed a good agreement with certain error. The procedure, however, predicts more conservative results in terms of the maximum equivalent plastic strain (PEEQ). In order to estimate the accuracy in the recently developed model, a series of nonlinear FEA pipe indentation simulations were conducted using the finite element analysis tool, ABAQUS and compared with the analytical prediction.

Recognizing the inherent error in the analytical model for dent strain assessment, machine learning techniques e.g., Gaussian Process Regression (GPR) and random forest (RF), was used for the accuracy assessment of the developed analytical model, quantifying the error in comparison with the FEA in terms of the maximum PEEQ. By varying the dent depth and the indenter radius, a model that quantifies the error inherent in the analytical model was developed. The proposed error model and the original analytical model along with the accuracy of the error prediction can be utilized to rapidly determine the severity of a dent.

PREFACE

This thesis is an original work by Yueying Li. Chapter 5 of this thesis has been previously published as:

Li, Y., Li, Y., Hassanien, S., Okoloekwe, C., & Adeeb, S. (2019) "Application of Gaussian Process Regression for the Accuracy Assessment of a Three-Dimensional Strain-Based Model." Proceedings of the ASME 2019 Pressure Vessels & Piping Conference. San Antonio, Texas, USA.

For the above work, I was responsible for the data collection, analysis, and manuscript composition. Okoloekwe. C assisted with the validating the models. Li. Y and S. Adeeb assisted in concept formation and provided technical support, and the results of the investigation were reviewed by Hassanien, S.

ACKNOWLEDGEMENTS

I would first like to thank my supervisors, Dr. Adeeb and Dr. Li, for their continued support and guidance throughout my research.

I am also thankful to my family, friends, and partner, for their love and encouragement during my years of study. This achievement would not have been possible without them. I would especially like to thank my parents for always believing in me and supporting me.

TABLE OF CONTENTS

Table of Figures

List of Abbreviations

ASME	American Society of Mechanical Engineers
CSA	Canadian Standards Association
FEA	Finite Element Analysis
GP	Gaussian Process
R	Radius of the Indenter
RF	Random forest
ILI	In-Line Inspection
MOP	Maximum Operating Pressure
MSE	Mean Squared Error
MSP	Most Significant Point
NEB	National Energy Board
OD	Outer Diameter
Oob	out-of-bag
PHMSA	Pipeline and Hazardous Materials Safety Administration
SMYS	Specified Minimum Yield Strength

CHAPTER 1: INTRODUCTION

1.1 Background

Pipelines are the safest and most efficient way to transport oil and gas products throughout the world. Today, more than 840,000 km of pipelines are laid out across Canada, beginning in Alberta and going west to British Columbia, north to the Northwest Territories, south to Texas, and east to Quebec. Gathering pipelines, feeder pipelines, and transmission pipelines transport crude oil, natural gas, and liquefied natural gas from wells to collection points, across provincial and national borders. Distribution pipelines deliver some natural gas products directly to consumers. According to the Canadian Association of Petroleum Producers, more than 99% of all the oil and gas moved by pipeline in Canada arrives at its destination (Reeves & Ryan, 2020).

Most pipelines are constructed of steel and formed by welding sections of pipe together. After the welds are X-rayed to detect any flaws, the pipe is wrapped with a protective coating and buried. The usual depth of burial is about 1.5 m for large pipes and slightly less for small pipelines. All pipelines, regardless of type, are then inspected and pressure tested before being put into service. Since pipelines are constructed with a combination of good design, materials, and operating practices, they do have a good safety record. However, like any other engineering structures, pipelines are subjected to different loading and environmental conditions, which can potentially cause pipeline failure. The most common threats of onshore and offshore, oil and gas transmission pipelines are mechanical damage (dents and gouges) and corrosion (Cosham and Hopkins, 2003). As the majority of pipelines are buried underground, those threats to the structural integrity of a pipeline can go undetected. If the gouges and cracks associated with rerounded dents are not discovered in time, severe consequences such as pipeline spill and leak will contaminate

local waterways, environment or even impact public health. As such, assessment methods are needed to determine the severity of such defects when they are detected in pipelines. This research will focus on dents.

After being in operation for years, many pipelines experience issues that cannot be detected from the outside, but these must be diagnosed to prevent further damages, potential revenue loss, and environmental concerns. Since not all pipelines can be easily inspected because they may be buried underground or be laid on a seabed, in-line Inspection (ILI) tools, sometimes referred to as "smart pigs", are thus commonly used to inspect pipelines for evidence of internal corrosion, deformations, laminations, cracks, or other defects. ILI tools utilize caliper or sensors to collect data of inner diameter of the pipe and indicate the location, shape, and size of dents. ILI tools run through pipelines and report the abnormal geometry of the pipeline. The data obtained from such inspections are used to detect the anomaly and assist decision-making on structural integrity of the pipeline. Guidance is therefore needed so that operators can identify dents that could potentially be damaging to pipeline integrity and need excavation.

Current industry regulations recommend repairing of dents primarily based on depth by setting acceptability limits. A summary of the code and best practice guidance for the assessment of dents in pipelines is provided in Table 1.

Table 1 Acceptability limits for plain dents (Race, et al., 2010)

Stipulation	Plain dents	
	Constrained	Unconstrained
ASME B31.8	Up to 6% OD or strain level up to 6%	
ASME B31.4	Up to 6% OD in pipe diameters > NPS4" Up to 6 mm in pipe diameters < NPS 4"	
API 1156	Up to 6% OD but > 2% OD requires a fatigue assessment	
EPRG	≤ 7% OD at a hoop stress of 72% SMYS	
PDAM	Up to 10% OD	Up to 7% OD
CSA Z662	Up to 6 mm for ≤ 101.61 mm OD, or < 6% OD for > 101.6 mm OD	

Although the dent depth is one of the most symbolic characterization of a dent, it has been noted that the depth of the dent is not sufficient to take as the only principle to check if the dent area needs to be repaired. Many pipelines failed at dent locations but they fell within codified criteria. For example, it can be caused by fatigue. In other situations, many less severe dents are excavated according to relevant regulations. Therefore, there is a requirement for a better guidance to ensure the operation safety and economy of pipelines.

The failure from practice indicates that the severity of a dent cannot be fully assessed based on its depth alone. The restraint condition, indenter shape, loading sequence, and pressure-cycling history should also be considered when assessing dents (Kainat et al., 2019). Although the ASME B31.8 (The American Society of Mechanical Engineers, 2010) provides the non-mandatory analytical strain to predict the equivalent strain on the surface of the dented pipeline, there is currently no codified methodology for conducting a dent fatigue assessment.

As mentioned before, there is no general equation or model to assess the severity of a dent as the severity of a dent is affected by too many different variables (including size, shape, interacting features, and the pipe's operating conditions). In order to compensate for the lack of versatility and accuracy of the codified techniques, numerical modelling via finite element analysis

(FEA) has been proposed in literature as an accurate technique to evaluate dents (Arumugam, Gao, Krishnamurthy, Wang, & Kania, 2016; Hassanien, Kainat, Adeeb, & Langer, 2016). The nonlinear numerical modelling with FEA is used to model the full geometry of dents and simulate the indentation of the pipeline reported by inline inspection (ILI). As such, the most accurate assessment of the stress-strain conditions associated with the deformation of a variety of materials with complex geometries under complex loading scenarios can be provided by using FEA. The issue facing pipeline operators is that a large number of dents are being reported by ILI tools and it is impossible to conduct FEA on every dent.

To precisely match with the reported dent geometric profile, FEA can be computationally demanding and expensive as it can require numerous simulation trials to be run. In order to handle this complex problem with a large number of inputs, Okoloekwe (2017) developed a mathematical approach to interpolate a dented surface using spline functions to evaluate the radius of curvature, and subsequently the strain, at any point on a dented region of a pipeline. Compared to FEA, the analytical technique provided a relatively good prediction of strain distribution and the algorithm was simple to implement. However, the analytical method consistently predicted conservative strain values compared to nonlinear FEA for all the models investigated.

This study aims at further advancing the proposed technique by quantifying the prediction error in the maximum equivalent plastic strain (PEEQ) benchmarked against the values by nonlinear FEA. Specifically, this study used Gaussian Process (GP) modelling technique to learn the difference between the mathematical model and the nonlinear FEA in terms of the maximum PEEQ.

A machine-learning algorithm that involves a Gaussian process uses lazy learning and a measure of the similarity between points (the kernel function) to predict the value for an unseen

point from training data. The prediction is not just an estimate for that point, but also has uncertainty information. The database of samples used to train the neural network can be based on numerical analyses.

1.2 Study Objectives

1.2.1 Main Objective

The main objective of this thesis is to work towards an improved assessment methodology of the integrity of dented pipe. The research evaluates the use of finite element analysis to model dents reported by in-line inspection tools and assess their severity. In addition, the research considers the use of Gaussian process (GP) to predict the stress-strain state of dents using results from FEA and ultimately bypass the need to develop the time-intensive models.

1.2.2 Specific Tasks

In order to achieve the main objective of this study, a number of specific tasks were defined and sequentially pursued. They include:

The first task was to develop a procedure to model indentations in pipelines using FEA, which includes determining specific properties of the model that would lead to obtaining consistent and accurate results in an efficient manner.

The second task is to describe how to use FEA automatically to create an efficient dent analysis process and the deformed geometry of the dented surface of the pipe are extracted and interpolated with B-spline curves

The third task is to use Gaussian Process and propose the random forest techniques to provide tools for estimating dent severity while bypassing the need for computationally demanding finite element analysis models.

1.3 Organization of Thesis

This thesis is organized into seven chapters:

Chapter 1 describes the background of pipeline integrity and current methods for dent assessment on pipelines.

Chapter 2 is a comprehensive literature review of historical dent failures and existing methods for dent assessment, which includes analytical and numerical methods. Proposed finite element analysis procedures for modelling dents will also be reviewed, as well as areas requiring additional research from the current methods. Literature describing background of Gaussian process regression and random forest will also be discussed.

Chapter 3 presents the general FEA methodology that will be used throughout this research, as well as the methodology and results of the mesh configuration studies. This includes the identification of an optimal mesh size in the indented region, an applicable element type, and the number of thickness integration points that should be used in this research for accuracy and efficiency. In addition, the procedure will be validated against experimental results.

Chapter 4 reproduces Okoloekwe 's three-dimensional analytical model and the prediction are benchmarked against the predictions of a numerical model and the codified dent strain expressions.

Chapter 5 gives a basic introduction to Gaussian Process regression models by focusing on understanding the role of the stochastic process and how it is used to define a distribution over functions. The simple equations are presented for incorporating training data and examine how to learn the hyperparameters using the marginal likelihood.

Chapter 6 presents the notion of an ensemble learning methods, bagging, and describes a method of building a forest of uncorrelated trees using a CART like procedure, combined with

randomized node optimization and bagging. The estimation of the performance of each model on its left out using out-of-bag error and measuring variable importance through permutation will also be discussed. This serves as a proposal for potential work using random forest for dent analysis of pipelines.

Chapter 7 summarizes the key findings through this research and recommends the direction for future research work.

CHAPTER 2: LITERATURE REVIEW

A review of the literature was conducted to study how current guidelines and the previous research works address the significance of dent in pipeline. It was found that the dent depth as a percentage of outer diameter of the pipe, which is a geometric parameter, is most commonly used by different codes, standards, and manuals for determining the severity of a dent. Some research works reported in the literature focused on the concentration of strain in a pressurized dent. Though current codes, standards, and manuals consider depth as the only geometric parameter for assessing the severity of the dent, previous research works indicated that use of depth alone may results an underestimation or overestimation of dent severity. Although FEA has proven accurate in providing the stresses and strains within a dented region, FEA is inefficient for analyzing a large number of dents as it is computationally expensive. Consequently, Okoloekwe (2017) proposed a modification to the use of codified equations and developed a mathematical approach for evaluating the strains in dented pipelines. In order to accuracy assessment of the developed analytical dent strain assessment model and for quantifying the error in comparison with the FEA in terms of the maximum equivalent plastic strain, we turned to Gaussian Processes Regression (GPR), a powerful, non-parametric regression technique with solid probabilistic foundations. The main advantages of GPR over other approaches is that they provide well defined confidence intervals, which are very important to assess the quality of the model.

2.1 Introduction of Dents and Definitions

The study of dent has been ongoing for many years. Dents in pipeline are permanent plastic deformation of the circular cross section of the pipe (Cosham and Hopkins 2003). A dent is a gross distortion of the pipe cross-section.

A photograph of a dent in pipeline is shown in ***Figure 1***.

Figure 1 Photograph of a dent in the pipe wall (Source:http://www.google.ca/images)

Dent depth (H) is defined as the difference between the maximum reduction in diameter and the original diameter of the pipe (Cosham and Hopkins 2003) (see **Figure *2***). This depth includes local indentation and any divergence in the overall cross-section (Cosham & Hopkins, 2004).

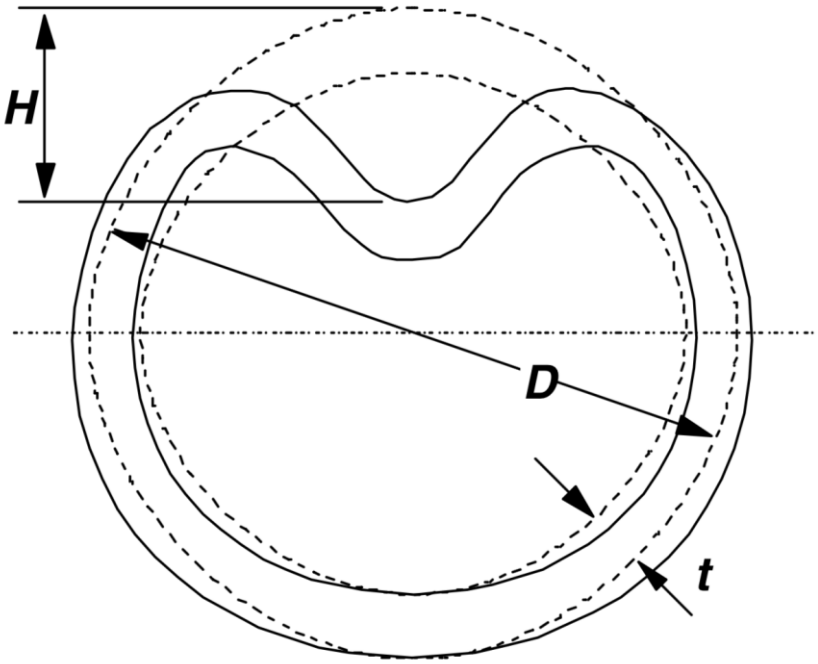


Figure 2 The dimensions of a dent (Macdonald et al. 2006)

Dents are often classified into different categories. Based on the curvature of the dent it can be classified as smooth dent and kinked dent. A smooth dent is one which causes a smooth change in the curvature of the pipe wall. A kinked dent causes an abrupt change in the curvature of the pipe wall (Cosham and Hopkins 2003). However, there is no universally accepted value of the threshold curvature that differentiates the two dents. According to European Pipeline Research Group (EPRG), a dent can be classified as a kinked dent when the radius of curvature (in any direction) of the sharpest part of the dent is less than five times the wall thickness of the pipe (Roovers et al. 2000). Photograph of kinked dent is presented in **Figure 3**.

Figure 3 Photograph of kinked dent in field pipeline (Macdonald et.al. 2006)

Depending on the surrounding conditions and constraints, dents can be classified as constrained dent or as unconstrained dent. A constrained dent is the one which is not free to rebound, because the indenter (such as a rock) remains in contact with the pipe and prevents the dented region from further movement (Cosham & Hopkins, 2004). A dent which is free to rebound elastically when the indenter is removed and is free to reround with the increasing internal pressure is termed as an unconstrained dent (Cosham and Hopkins 2003).

Dents are generally classified as a plain dent or dent with associated mechanical damages. These damages include gouges, girth welds and any damage with mechanical removal or displacement of metal that causes a measurable reduction in thickness (Cosham and Hopkins 2003). Instead, a plain dent is defined as a smooth dent (a dent which causes a smooth change in the curvature of the pipe wall) without reductions of the wall thickness or other defects (Andrew &

Phil, 2003). Plain dent is often found in the field pipelines and it is the main focus of the current research project.

Dents in pipelines are usually caused by the third-party damage or construction damage. Third party damage generally occurs on the upper half of the pipe and the corresponding dents are constrained. Whereas, construction-caused dents normally appear on the bottom half of the pipe and tend to be constrained, i.e. a rock beneath the pipeline (Race et al., 2010). All dents have the potential to induce localized stress and strain concentration and then result in pipeline failures. Thus, it is necessary to make a consistent assessment to investigate the effect of dents on pipeline integrity.

Although the dent depth is the most significant factor indicate the severity of a plain dent, the stress and strain distributions vary based on the dent length and width. Studies have shown that the maximum stress and strain occurs at the root (See **Figure 4**) of the long dents, whereas at the flanks (See **Figure 4**) of the short dents (Cosham & Hopkins, 2004).

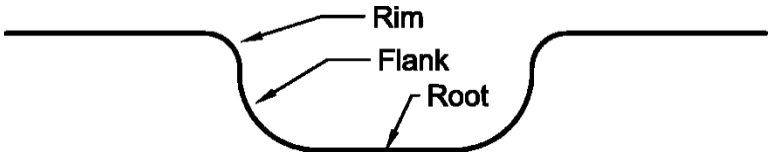


Figure 4 Nomenclature of Typical Dent (Oshana Jajo, 2014)

Overall, pipelines are widely used in the industry to carry and transport oil or pressurized gas. Having a dent defect in a pipeline introduces strain and stress concentrations that must be examined in order to determine the structural integrity and safety of the line pipe. The determination of strains in dents has been a major topic for research for a long period of time. Different methods were developed to determine the strains in different type of dents as there is no specific method of determining strains for all types of dents. It was found from the literature review that plain dents, which are the focus of this research project, are the least dangerous types of dents encountered in pipelines.

2.2 Effect of Dents on Pipeline Integrity

In 2015, a gasoline leak of a petroleum pipeline was detected in Centreville, VA. During the initial excavation, there was no rock or other object in direct contact with the pipeline. Reports stated that a rock had been removed in a prior excavation in 1994. This dent depth was calculated as being 2.3% of the pipe diameter. After failure investigation, it was found that a dent on the underside of the pipe led to the formation of a crack, which then propagated until the penetrate the wall due to pressure cycling and the pipe leaked.

As of 2019, there were over 224,000 miles of hazardous liquid pipeline in the USA and Canada, of which over 63,000 miles carry refined petroleum products (PHMSA, 2019).

Steel pipelines must be designed to resist a variety of conditions. Because most of these pipelines lie underground, which in turn increase the risk for the creation and detection of damages. One of the issues studied are dents in pipelines.

Unconstrained dents are formed when the object that indented the pipe is removed sometime afterwards typically happen at the top half of a pipeline. Rock dents (constrained dents) that have the rock remain in the same position are found at the bottom of a pipeline. And if the rock is sharp enough and the internal pressure of pipeline is high, it could lead to puncture failure. Mechanical damages in the form of dents could result in delayed failure although the damage is not severe enough to cause immediate failure. Dents may be associated with coating damage and allow the pipe metal surface expose to oxygen, moisture and other contaminants, and hence may be sites for the initiation of corrosion cracking. In addition, dents can create stress risers in the pipe, allowing initiation sites for fatigue and stress corrosion cracking.

In most studies, the analysis of mechanical damages was mainly through experimental or analytical studies as well as Finite Element method. And dent depth was used as the only parameter

for the determination of pipeline safety in these studies as well as existing standards and codes (Rosenfeld, 2002). In addition, other studies used the strain-based criteria for the evaluation of the severity of dents (Lancaster and Palmer 1996; Rafi et al. 2012). All methods are discussed in the following sections.

2.3 Existing Dent Assessment Methods

In previous studies, the strains in dented pipelines was determined in two ways: when dent was being formed and when the dented pipe was being pressurized. The severity of dent can be obtained by determining the strain level of the dent. While extensive industry efforts have focused on a suitable integrity assessment method for pipeline dents, a single, agreed-upon analytical or empirical model does not currently exist (Hassanien et al., 2016). Strains in the dented region of the pipe can be obtained either numerically using finite element (FE) method. Determining strain values using FE method requires the solution of large plastic deformation shell with large number of nodes and large amount of time for each dent assessment and thus are not suitable for system-wide application (Hassanien & Langer, 2018).

A comparison between the results obtained from the finite element analysis (FEA) and the experimental or analytical results is usually carried out to validate the numerical (FE) model.

2.3.1 Regulatory Requirements

The current regulatory requirements for repairing mechanical damage in both liquid and gas pipelines rely on reported inspection (ILI) measurements of deformation features: the nature of the mechanical damage (i.e. plain dents, dents with gouges, cracks or welds, etc.) and dent depth.

The Federal Regulations for liquid and gas pipelines in the United States set the acceptable limit for plane dents as follows:

- A dent that has any indication of metal loss, cracking, or a stress riser
- A dent with depth >6% of the nominal outer diameter (OD) of the pipe
- A dent with depth >3% OD on the upper 1/3 of the pipe
- A dent with depth >2% OD on the upper 2/3 of the pipe
- A dent with depth >2% OD that affects the pipe curvature at a girth weld or longitudinal weld

These requirements are based on the assumption that dents on the top of the pipeline are more likely to have been caused by third party damage, instead the rock dents are more likely occur on the bottom of the pipeline. Mechanical damage is one of the most severe forms of pipeline defect as it is often accompanied by cracking and gouging and thus lead to low burst pressures and fatigue lives. Consequently, the remedial measure for upper dents are more urgent than for rock dents.

Canadian pipeline code, CSA Z662, proposes the dents that meet the following criteria must be repaired:

- A dent associated with a crack or gouge
- A dent with depth >6% OD
- A dent with depth >2% OD at a weld
- A dent with a corrosion feature that does not meet the criteria described in ASME B31.G (The American Society of Mechanical Engineers, 2012)

2.3.2 Strain-Based Methods

Literature has shown that strain-based criteria are more accurate than existing depth-based criteria and being adopted by more and more associations. The use of strain-based criteria better accounts for the localized distortion of the dent area and the curvature of the profile of the dent.

The ASME B31.8-2003 provides the non-mandatory strain equations to use data provided by ILI tools to estimate the strains in dents. The equations use the radius of curvature, length, and

depth as input variables. However, how to implement these formulas is not provided (as to how to obtain the variables) and as such its use has to depend on the expertise of the operator.

Limitations of the ASME B31.8 strain equations result from that the ASME B31.8 equations based on the assumption that the maximum strain coincidently occurs at the dent apex. But as the location of the peak strains are highly dependent on the geometry of the dent, it does not apply for all cases. The circumferential membrane strains and shear strain components have however been ignored in this computation as they are believed to have a little effect on the global strain state of the pipeline. However, there might be some deviation from this assumption according to FEA models of real-life dents.

Additional analytical strain estimation methods have been proposed that intended to improve on the aforementioned shortcomings of the ASME B31.8 equations (Lukasiewicz et al., 2006; Noronha et al., 2010; Gao & Krishnamurthy, 2015; Okoloekwe, 2017). The methods were validated using FEA and were proven to produce similar results to FEA while requiring far less time and resources. Okoloekwe (2017) presented a method of interpolating a dented surface using spline functions to assess the radius of curvature, and subsequently the strain, at any point on a dented section of a pipeline. Compared to FEA, the analytical methodology presented by Okoloekwe (2017) proved accurate, but conservative. Although quicker and easier to perform than FEA, ASME B31.8 equations, and other similar analytical strain equations, cannot be used to predict the strain of dents interacting with stress risers or dents with complex shapes, such as dents with multiple apexes.

2.3.3 Fatigue-Based Methods

Assessment based upon either depth or strain alone may not provide non-conservative results because unlike metal loss, a dent does not simply fail due to the operating pressure exceed its capacity or leak by through-wall corrosion. A dent can induce the initiation of a crack, which may propagate with the subsequent re-rounding and the operational pressure cycling, exhausting the material's fatigue life (Turnquist & Smith, 2016). Once a crack is formed in a dent it will grow quickly due to the stress concentration until the crack reach through-wall and the pipe leaks.

From the research presented by Turnquist and Smith (2016), the remaining life of a dent is estimated based on the stress concentration and the effect of pressure cycling at the dent location. Stress concentration factors are calculated based on the peak hoop and axial stresses using FEA. Past representative pressure cycling is then processed using a rainflow counting approach. The amount of damage accumulated during each pressure cycle is calculated using stress or strain-based fatigue curves and can be converted to a damage rate to estimate the dent's remaining life.

2.4 Finite Element Analysis

The Finite Element Analysis (FEA) is a numerical method for solving problems with complicated geometries, loadings, and material properties where analytical solutions cannot be obtained by discretizing them into smaller, simpler parts that are called finite elements.

FEA has been used for many studies of pipeline dents and has been verified by the full-scale denting tests presented in the literature. While FEA is a time- consuming process involving detailed analysis of each feature, it provides the most dependable and accurate assessment of the stresses and strains to account for a variety of material properties, complex loading situations, and interaction with stress concentrators (Arumugam et al., 2016; Ghaednia & Das, 2018; Hassanien et al., 2016; Pinheiro et al., 2014).

There are several FE software packages such as Ansys, Abaqus, Nastran, Patran and so on. Though the software interfaces are different, they all follow the same procedure in solving problems. This section describes the background of several aspects of FEA and how it is used by other researchers.

2.4.1 Material Properties

The uniaxial true stress versus true strain curve for the appropriate steel grade being modelled is sufficient for use in the FEA models of pipe steels, according to Hyde, Luo, and Becker (2011) and Arumugam et al. (2016). Validated models have confirmed that it is sufficient to assume that all materials obey an isotropic hardening rule (Hyde et al., 2011; Kainat et al., 2019).

2.4.2 Indenter Properties

The indenter has been modelled a rigid surface and the interaction between the pipe surface was modeled as “hard contact”. The indentation was modeled by applying a vertical downward displacement on the rigid indenter, which created the pipe deformation.

2.4.3 Boundary Conditions

The modelling of only a half, or quarter of the pipe segment and utilizing symmetry boundary conditions has been proposed in literature to reduce computational efforts (Hyde et al., 2011; Tiku et al., 2012; Arumugam et al., 2016). In several procedures, the bottom edge of the pipe (opposite from the indenter) is restrained from movement in the vertical direction (Tiku et al., 2012; Arumugam et al., 2016).

2.4.4 Mesh Configuration

There are various mesh configurations for FEA models used in literature. Many sources use a fine mesh in the indentation region and coarser mesh elsewhere to ensure the most accurate results are obtained in the indentation region but to not sacrifice computational time where

accuracy is not needed (away from the indenter) (Arumugam et al., 2016; Hassanien et al., 2016; Hyde et al., 2011). A partition can be created in the indentation region to allow this mesh configuration to be possible (Hyde et al., 2011).

2.5 Gaussian Process

The GP model is a probabilistic, non-parametric model, which differs from most of the other parametric models by fitting the parameters of the selected basis functions (Williams, 1998). By contrast, GP searches for the relationship from the observed data.

The GP approach finds a distribution over the possible functions $f(x)$ that are consistent with the observed data. As with all Bayesian methods, it begins with a prior distribution and updates this as data points are observed, producing the posterior distribution over functions. The covariance matrix can smooth the function we are interested in by ensuring that values that are close in the input space will produce output values that are close. This covariance matrix, along with a mean function, defines a Gaussian Process.

Gaussian processes can be used as a supervised learning technique for classification as well as regression. An example of a classification task would be to recognize handwritten digits, whereas an example of a regression problem would be to learn the inverse dynamics of a robot arm (Bocsi et al., 2011). For the latter, the task is to obtain a mapping from the state of the arm (given by the positions, velocities and accelerations of the joints) to the corresponding torques on the joints. Such a mapping can then be used to compute the torques needed to move the arm along a given trajectory. Another example is predictive soil mapping (Gonzalez et al., 2007), where one is given a set of soil samples taken from some regions and asked to predict the nature of soil in another region. A major benefit of using Gaussian processes to solve these problems is that they can provide confidence measures for the predictions. For instance, in the context of predictive soil mapping, one can use Gaussian processes to decide which regions should be given a higher priority for collecting soil samples, based on the uncertainty of the predictions. The following sections

provide a mathematical treatment of Gaussian processes, and their application to regression problems.

2.5.1 Gaussian Process

A Gaussian process can be thought of as a Gaussian distribution over functions (thinking of functions as infinitely long vectors containing the value of the function at every input). Formally let the input space be $\mathcal{X}$ and $f : \mathcal{X} \to \mathbb{R}$ be a function from the input space to the reals. We then say that f is a Gaussian process if for any vector of inputs $x = [x_1, x_2, \dots, x_n]^T$ such that $x_i \in \mathcal{X}$ for all i, the vector of outputs $f(x) = [f(x_1),\ f(x_2), \dots, f(x_n)]^T$ is Gaussian distributed.

A Gaussian process is specified by a mean function $\mu : \mathcal{X} \to \mathbb{R}$, such that $\mu(x)$ is the mean of $f(x)$ and a covariance/kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $k(x_i, x_j)$ is the covariance between $f(x_i)$ and $f(x_j)$. We say $f \sim GP(\mu, k)$ if for any $x_1, x_2, \dots x_n \in \mathcal{X}$, $[f(x_1), f(x_2), \dots, f(x_n)]^T$ is Gaussian distributed with mean $[\mu(x_1), \mu(x_2), \dots, \mu(x_n)]^T$ and $n \times n$ covariance/kernel matrix K_{xx}:

$$K_{xx} = \begin{bmatrix} k(x_1,x_1) & k(x_1,x_2) & \cdots & k(x_1,x_n) \\ k(x_2,x_1) & k(x_2,x_2) & \cdots & k(x_2,x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n,x_1) & k(x_n,x_2) & \cdots & k(x_n,x_n) \end{bmatrix} \tag{2.1}$$

The kernel function must have the following properties:

- Be symmetric. That is, $k(x_i, x_j) = k(x_j, x_i)$
- Be positive definite. That is, the kernel matrix K_{xx} induced by k for any set of inputs is a positive definite matrix.

Examples of some kernel functions are given below:

• Squared Exponential Kernel (Gaussian/RBF): $k(x_i, x_j) = e^{-\frac{1}{2\gamma^2}(x_i - x_j)^2}$ where γ is the length scale of the kernel.

• Laplace Kernel: $k(x_i, x_j) = I(x_i = x_j)$, where I is the indicator function.

• Linear Kernel: $k(x_i, x_j) = x_i^T x_j$.

More complicated kernels can be constructed by adding known kernel functions together, as the sum of two kernel functions is also a kernel function.

A Gaussian process is a random stochastic process where correlation is introduced between neigh- boring samples (think of a stochastic process as a sequence of random variables). The covariance matrix K_{xx} has larger values, for points that are closer to each other, and smaller values for points further apart. This is illustrated in **Figure 5**. The thicker the line, the larger the values. This is because the points are correlated by the difference in their means and their variances. If they are highly correlated, then their means are almost same, and their covariance is high.

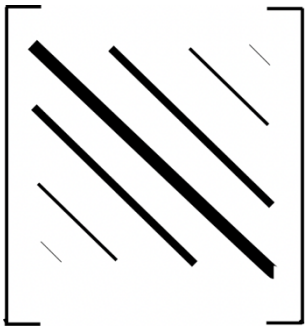


Figure 5 Magnitude of values in covariance matrix

2.5.2 Visualizing Samples from a Gaussian Process

To actually plot samples from a Gaussian process, one can adopt the following procedure:

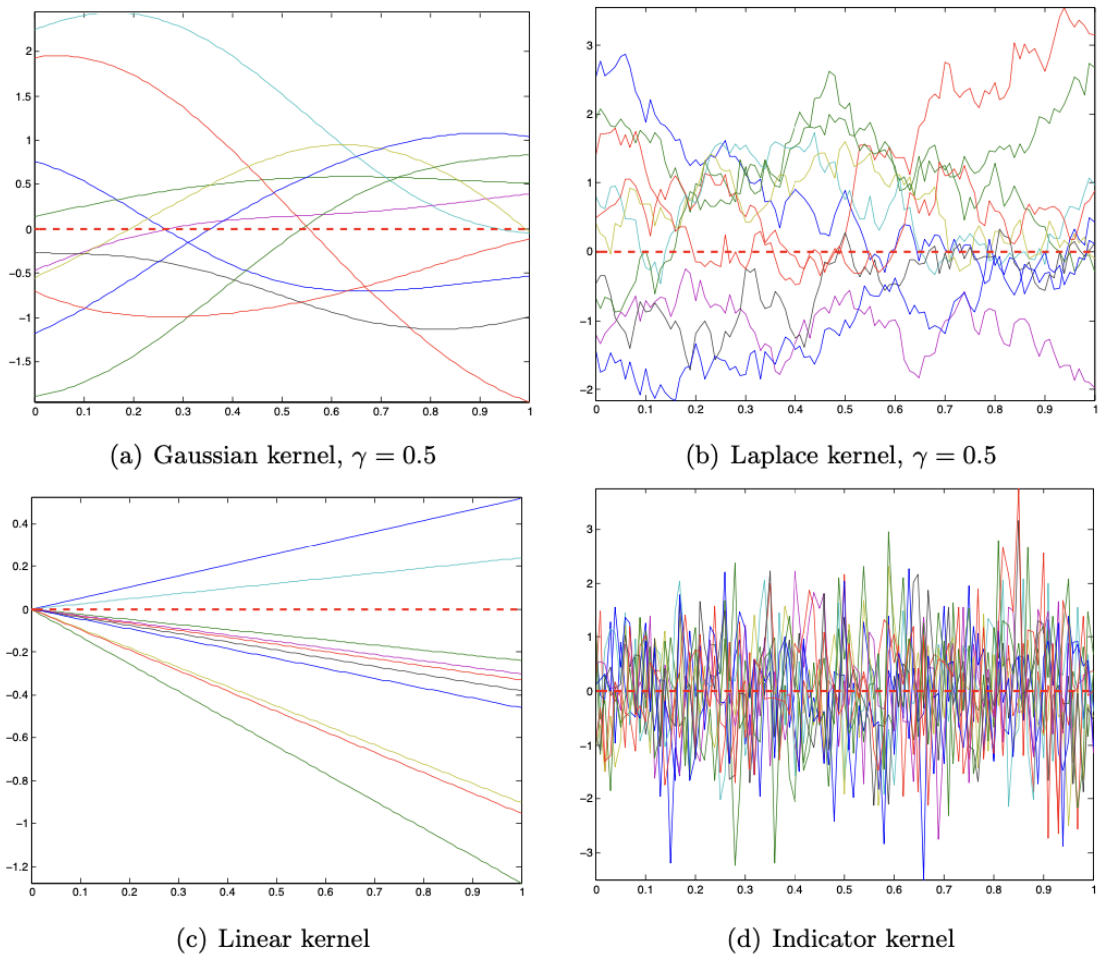
1. Define the mean function and kernel as prior. For instance, $\mu = 0$, and $k(x_i, x_j) = e^{-\frac{1}{2\gamma^2}(x_i - x_j)^2}$, with $\gamma = 0.5$.

2. Sample from prior, example: $x_i = \varepsilon \times i,\ i = 0,1, \ldots, \frac{1}{\varepsilon}$

3. Compute the kernel matrix Σ. For the example with $\varepsilon = 0.1$, we would have a 11×11 matrix.

4. Predict from the multivariate Gaussian distribution $N(0, \Sigma)$

5. Plot the samples



(a) Gaussian kernel, $\gamma = 0.5$ (b) Laplace kernel, $\gamma = 0.5$

(c) Linear kernel (d) Indicator kernel

Figure 6 Samples from a Gaussian process

Figure *6* shows samples of different kernels with $\varepsilon = 0.01$ drawn from a Gaussian process.

2.5.3 Inference

Gaussian processes are useful as priors over functions for doing non-linear regression. In **Figure 7(a)**, we see a number of sample functions drawn at random from a prior distribution over

functions specified by a particular Gaussian process, which favours smooth functions. This prior is taken to represent our prior belief over the kinds of functions we expect to observe, before seeing any data.

Note that $\frac{1}{N}\sum_{i=1}^{N} f_1(x) \rightarrow \mu_f(x) = 0$,as $N \rightarrow \infty$

Now, given a set of observed inputs and corresponding output values $(x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_n, f(x_n))$, and a Gaussian process prior on f, $f \sim GP(\mu, k)$, we would like to compute the posterior over the value $f(x^*)$ at any query input x^*. **Figure *7(b)*** shows sample functions drawn from the posterior, given some observed (x, y) pairs. We see that the sample functions from the posterior pass close to the observed values but vary a lot in regions where there are no observations. This shows that uncertainty is reduced near the observed values.

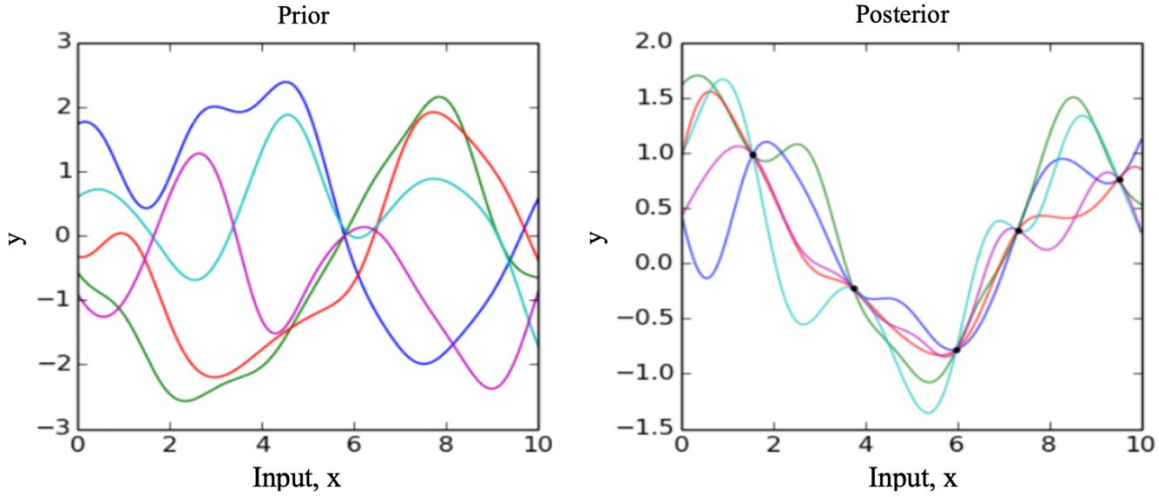


Figure 7 Gaussian process inference

2.5.3.1 Computing the Posterior

The posterior can be derived similarly to how the update equations for the Kalman filter were de- rived. First, we will find the joint distribution of $[f(x^*), f(x_1), f(x_2), \ldots, f(x_n)]^T$, and then use the conditioning rules for a Gaussian to compute the conditional distribution of $f(x^*)|f(x_1), \ldots, f(x_n)$.

Assume for now that the prior mean function $\mu = 0$. By definition of the Gaussian process, the joint distribution $[f(x^*), f(x_1), f(x_2), \ldots, f(x_n)]^T$ is a Gaussian:

$$\begin{bmatrix} f(x^*) \\ f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} k(x^*, x^*) & k(x^*, \boldsymbol{x})^T \\ k(x^*, \boldsymbol{x}) & K_{xx} \end{bmatrix} \right) \tag{2.2}$$

where K_{xx} is the kernel matrix defined previously, and

$$k(x^*, \boldsymbol{x}) = \begin{bmatrix} k(x^*, x_1) \\ k(x^*, x_2) \\ \vdots \\ k(x^*, x_n) \end{bmatrix} \tag{2.3}$$

Using the conditioning rules we derived for a Gaussian, the posterior for $f(x^*)$ is:

$$f(x^*)|f(\boldsymbol{x}) \sim N(k(x^*, \boldsymbol{x})^T K_{XX}^{-1} f(\boldsymbol{x}), k(x^*, x^*) + k(x^*, \boldsymbol{x})^T K_{XX}^{-1} k(x^*, \boldsymbol{x})) \tag{2.4}$$

The posterior mean $\mathbb{E}(f(x^*)|f(\boldsymbol{x}))$ can be interpreted in two ways. We could group the last two terms $K_{XX}^{-1} f(\boldsymbol{x})$ together and represent the posterior mean as linear combination of the kernel function values:

$$\mathbb{E}(f(x^*)|f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i k(x^*, x_i) \tag{2.5}$$

for $\alpha = K_{XX}^{-1} f(\boldsymbol{x})$. This means we can compute the mean without explicitly inverting K, by solving $K\alpha = f(x)$ instead. Similarly, by grouping the first two terms $k(x^*, \boldsymbol{x})^T K_{XX}^{-1}$, the posterior mean can be represented as a linear combination of the observed function values:

$$\mathbb{E}(f(x^*)|f(\boldsymbol{x}) = \sum_{i=1}^{n} \beta_i f(x_i) \tag{2.6}$$

for $\beta = k(x^*, \boldsymbol{x})^T K^{-1}$.

2.5.3.2 Non-zero mean prior

If the prior mean function is non-zero, we can still use the previous derivation by noting that if $f \sim GP(\mu, k)$, then the function $f' = f - \mu$ is a zero-mean Gaussian process $f' \sim$

$GP(0, k)$. Hence, if we have observations from the values of f, we can subtract the prior mean function values to get observations of f', do the inference on f', and finally once we obtain the posterior on $f'(x^*)$, we can simply add back the prior mean $\mu(x^*)$ to the posterior mean, to obtain the posterior on f.

2.5.3.3 Noise in observed values

If instead of having noise-free observations of f, we observe $y(x) = f(x) + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$ is some zero-mean Gaussian noise, then the joint distribution of $[f(x^*), y(x_1), y(x_2), \ldots, y(x_n)]^T$ is also Gaussian. Hence, we can apply a similar derivation to compute the posterior of $f(x^*)$. Specifically, if the prior mean function $\mu = 0$, we have that:

$$\begin{bmatrix} f(x^*) \\ y(x_1) \\ \vdots \\ y(x_n) \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} k(x^*, x^*) + \sigma^2 & k(x^*, \boldsymbol{x})^T \\ k(x^*, \boldsymbol{x}) & K_{xx} + \sigma^2 I \end{bmatrix} \right) \tag{2.7}$$

The only difference with respect to the noise-free case is that the covariance matrix of the joint now has an extra σ^2 term on its diagonal. This is because the noise is independent for different observations, and also independent of f (so no covariance between noise terms, and between f and ε). So, the posterior on $f(x^*)$ is:

$$f(x^*)|y(\boldsymbol{x}) \sim N\big(k(x^*, \boldsymbol{x})^T (K_{XX} + \sigma^2 I)^{-1} y(\boldsymbol{x}), k(x^*, x^*) + \sigma^2 + k(x^*, \boldsymbol{x})^T (K_{XX} + \sigma^2 I)^{-1} k(x^*, \boldsymbol{x})\big) \tag{2.8}$$

2.5.3.4 Choosing kernel length scale and noise variance parameters

The kernel length scale (γ) and noise variance (σ^2) parameters are chosen such that they maximize the log likelihood of the observed data. Assuming a Gaussian kernel, we obtain the most likely parameters γ and σ by solving:

$$\max_{\gamma,\sigma} \log P(y(x)|\gamma, \sigma) = \max_{\gamma,\sigma} \left(-\frac{1}{2} y(x)^T (K_{XX} + \sigma^2 I)^{-1} y(\boldsymbol{x}) - \frac{1}{2} \log\left(det(K_{XX} + \sigma^2 I)\right) - \frac{N}{2} \log(2\pi)\right)$$

(2.9)

Here, the determinant term will be small when K_{XX} is almost diagonal; thus, this maximization favors smoother kernels (larger γ). Additionally, σ^2 can be chosen to have a higher value to prevent overfitting, since larger values for σ mean we trust observations lesser.

2.5.3.5 Computational complexity

One drawback of Gaussian processes is that it scales very badly with the number of observations N. Solving for the coefficients α that define the posterior mean function requires $O(N^3)$ computations. Note that Bayesian Linear Regression (BLR), which can be seen as a special case of GP with the linear kernel, has complexity of only $O(d^3)$ to find the mean weight vector, for a d dimensional input space $\mathcal{X}$. Finally, to make a prediction at any point, Gaussian process requires $O(N\hat{d})$ (where $\hat{d}$ is the complexity of evaluating the kernel), while BLR only requires $O(d)$ computations.

2.6 Random Forest

Random forest (RF) utilizes classification and regression tree (CART) as a learning algorithms of decision trees.

2.6.1 Bagging

RF are a highly versatile ensemble of decision trees that performs well for linear and non-linear prediction by finding a balance between bias and variance (Breiman 2001; Segal 2003). This ensemble learning method is known as 'bagging' as it grows trees in which successive trees do not depend on earlier trees. Each tree is independently determined using a bootstrap sample of the data set and a simple majority vote is taken for final prediction (Liaw and Wiener 2002). This is illustrated in **Figure 8** : Build a classification tree on a separate bootstrap sample of the training data, i.e., a random sample with replacement, and then use majority vote.

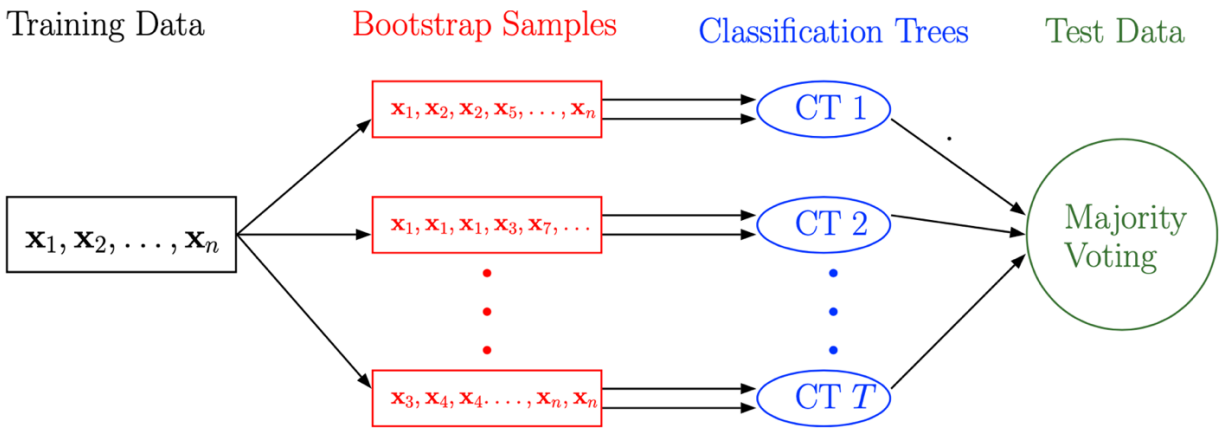


Figure 8 Building a random forest with bagging

2.6.1.1 out-of-bag

Drawing n out of n observations with replacement omits on average 36.8% of observations for each decision tree. We say that those samples left out by a tree are out-of-bag (oob) with respect to the tree. For each training example, we may vote the predictions of all the trees (for which the observation is oob) and compare with the true label to compute an average oob error. Such measure is an unbiased estimator of the true ensemble error and it does not require an independent validation dataset for evaluating the predictive power of the model.

The number of bootstrap samples/trees, T, is a free parameter. Typically, a few hundred to several thousand trees are used, depending on the size and nature of the training set. An optimal value of T can be found using cross-validation, or by observing the out-of-bag error.

2.6.2 From Bagging to Random Forest

The random forest approach is a bagging method where deep trees, fitted on bootstrap samples, are combined to produce an output with lower variance. However, random forests also use another trick to make the multiple fitted trees a bit less correlated with each other: when growing each tree, instead of only sampling over the observations in the dataset to generate a

bootstrap sample, we also sample over features and keep only a random subset of them to build the tree.

The motivation is to further de-correlate the trees in bagging: If one or a few features are very strong predictors for the class label, these features will be selected in many of the trees, causing them to become correlated.

Typically, the square root of the number of variables for classification and one third of the number of variables for regression are selected at random for each decision split.

CHAPTER 3: FEA MODELLING METHODOLOGY

3.1 Overview

Finite element model (FEM) was used to obtain the necessary information needed to achieve a better understanding of the experimental data and behaviour of dent subject to monotonically increasing internal pressure. This chapter emphasizes how the use of FEM was carried out to develop the numerical model. Validating the FE model with experimental test results is also necessary in order to have similar characteristics and to obtain comparable results with the experimental data. FE modeling was carried out to simulate the behavior of the experimental specimens by adopting similar geometry and material properties. The FEM tool ABAQUS version 6.14 distributed by SIMULIA was used to carry out the numerical modeling analysis. This tool was chosen as it is able to model pipelines with elasto-plastic isotropic and hardening material properties that are comparable to those from the experimental pipe specimens. Another reason to use such modeling tool is because it is one of the most popular and effective tools used to develop pipeline models that have a comparable denting load. This is demonstrated in other research such as the one conducted by Karamanos and Andreadakis (Karamanos and Andreadakis 2006). Furthermore, ABAQUS/Standard has also options for contact interaction that can simulate the experimental boundary conditions more precisely.

The purpose of generating a pipeline model with ABAQUS/Standard is to be able to more precisely predict the behaviour of a dented pipeline when it is being pressurized. Another reason is to obtain the strains within such dent in order to determine if a pipe is within the safe region. Lastly, a parametric study can also be conducted with the help of such modeling tool to develop a guideline that can be used to determine if a dented pipeline is safe. Such guideline will include

different types of indenters, internal pressures, and D/t ratios. With this finite element program, the pipe specimens were developed and tested under similar experimental conditions.

3.2 Modeling

First, a deformable, solid or shell, 3D part was sketched to represent a segment of the pipeline structure. The outer diameter and desired length of the pipe section were specified at this stage. An appropriate length of pipe section had to be selected that fully demonstrates the strains and stresses of the area of interest but was not too long that it was computationally exhaustive. For modelling dents according to ILI data, an appropriate indication to select the length of the pipe section is to use the entire length over which the inner diameter readings from the ILI tool seem to deviate from the nominal diameter.

Next, another part was created, separate from the pipe part, to represent the indenter. The indenter part was a 3D, analytical rigid body. The indenter shape used in this study is a spherical indenter.

3.3 Mesh Sensitivity Analysis

There are various mesh configurations for FEA models used in literature. Many sources use a fine mesh in the indentation region and coarser mesh elsewhere to ensure the most accurate results are obtained in the indentation region but to not sacrifice computational time where accuracy is not needed (away from the indenter) (Arumugam et al., 2016; Hassanien et al., 2016; Hyde et al., 2011). A partition can be created in the indentation region to allow this mesh configuration to be possible (Hyde et al., 2011).

For the mesh, eight-node brick elements were used with linear geometric order. The size of the mesh within the partition was 5 mm and the mesh grew coarser (15 mm) outside of the

partition, far away from the indentation (the area of interest). The pipe geometry was meshed with five elements through the thickness of the pipe wall to allow for the smooth transition from the tensile to the compressive zones in the dented region.

The indenters, however, were modeled as spherical rigid surface with various diameters.

3.4 Pipe

Each pipe specimen used was 1100 mm long with an outer diameter of 280 mm and wall thickness of 8 mm. The meshing technique was free meshing with quad-dominated element shape and three-dimensional 8-node brick elements. The meshing selected was a uniform mesh of 5 mm x 5 mm.

A partition was created to allow the area of the pipe closest to the most significant point (MSP), or minimum point, of the dent to have a fine, structured mesh and the mesh softened further away from the dented region so as to reduce computation time, which is pictured in **Figure 9**. FEA methodology described in literature typically follows this mesh configuration (Hyde et al., 2011; Arumugam et al., 2016).

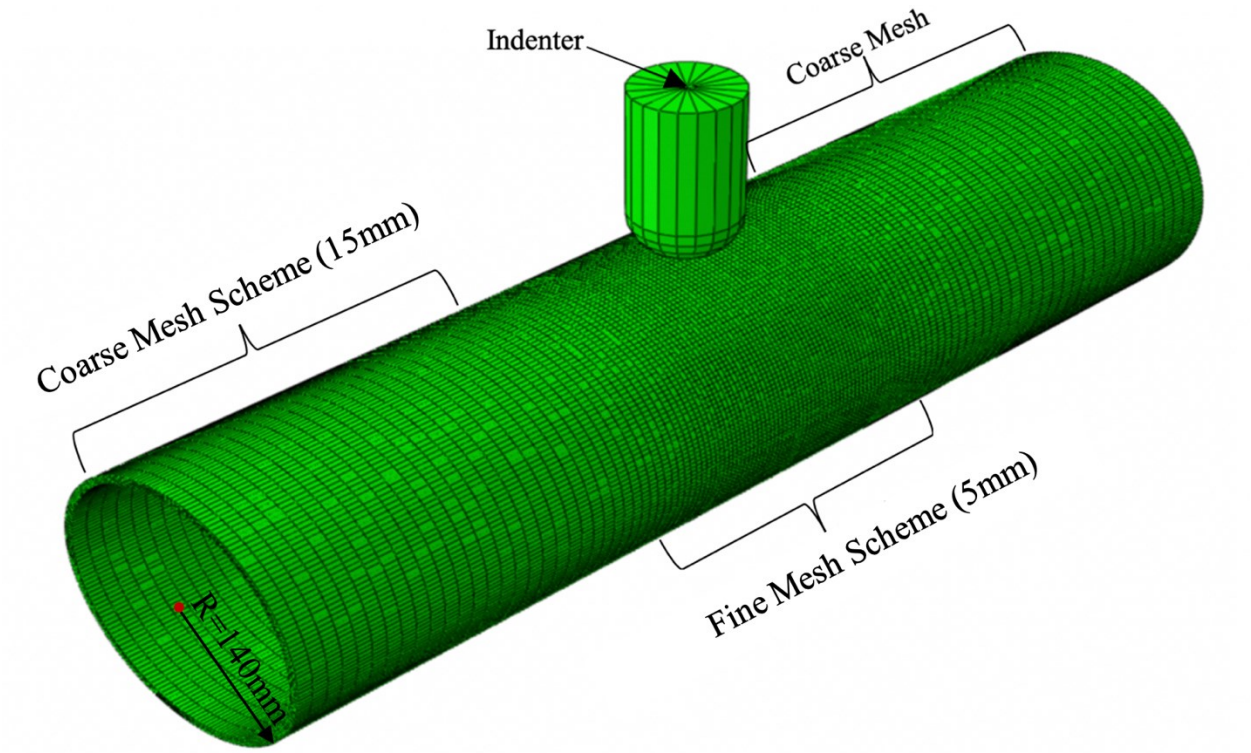


Figure 9 Finite element model of the pipeline

3.5 Material Properties

The material model for the pipeline is an isotopic elastoplastic material. The elastic behavior was assumed to be linear and isotropic with the elastic regime governed by a Young's Modulus of 200 GPa, Poisson's ratio of 0.3, while uniaxial stress strain curves defined the plastic behavior of the material. The plastic hardening was assumed to be isotropic. In CSA Z662 (2015), the stress-strain relationship can be estimated for steel pipe using Eq (3.1), shown below.

$$\varepsilon = \frac{\sigma}{E_S} + (0.005 - \frac{F_y}{E_S})(\frac{\sigma}{F_y})^n \tag{3.1}$$

where ε is the strain,

σ is the stress,

E_S is the modulus of elasticity of steel pipe,

F_y is the specified minimum yield strength,

n is the strain hardening parameter.

The pipe is an X-60 pipe having a yield stress, F_y, of 400 MPa.

The stress-strain relationship of a typical X-60 pipeline based on Ramberg-Osgood model(Eq (3.1)) is shown in **Figure 10**.

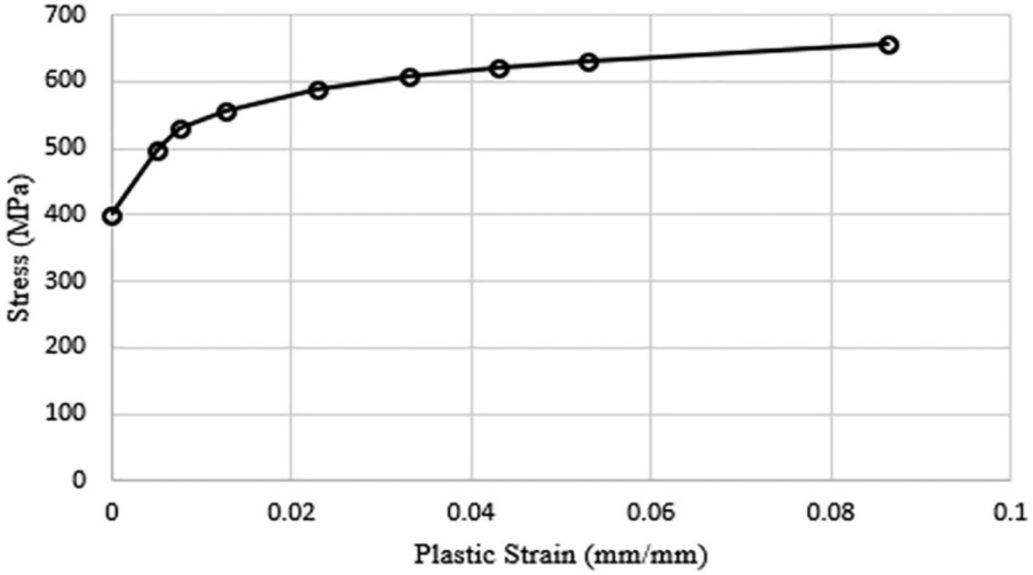


Figure 10 True stress-strain relationship of a typical X60 steel pipe (Okoloekwe, 2017)

3.6 Contact Formulation between Surfaces

In the denting tests (Rafi 2011), the indenters came into contact with the pipes, introducing an inward deformation on the pipe wall. In order to model this in ABAQUS, a contact algorithm was introduced. This contact acts between the indenter and the outer surface of the pipe. Additionally, a surface-to-surface (SS) discretization method was used for this FE model. This method is the standard discretization method provided in ABAQUS. With this approach, the surfaces having contact are required to be defined. A master surface and a slave surface must be defined in order to create a SS contact. ABAQUS provides guidelines regarding the selection of

such surfaces. The indenters were chosen as the master surface as their surface is stiffer than that of the pipelines. The pipeline was then chosen as the slave surface.

3.7 Boundary Conditions

The initial boundary conditions were set up such that a reference point at the center of both ends of the pipe was fixed and restricted from translations and rotations and restrained vertical translation at the bottom of the pipe. This was to simulate real-life pipe conditions that the support provided by the underlying soil prevents the vertical displacement of the buried pipelines. The boundary conditions are shown in **Figure 11**.

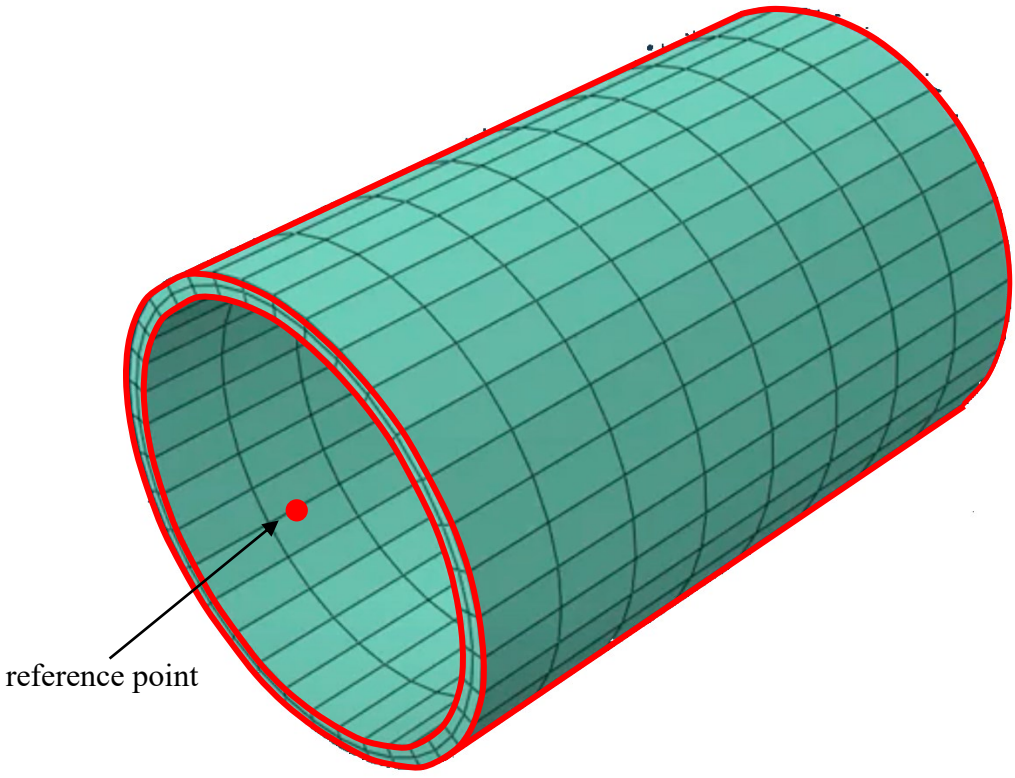


Figure 11 Boundary conditions for FEA pipe model

Oil and gas pipelines are constantly transporting fluid and the internal pressure was varied between 0% and 20% of the yield stress for different pipes specimens. While dents can occur anytime during its lifetime. The loading sequence chosen for the FEA study simulate the conditions experienced in the field. For constrained dent models, the indentation happened, and then a pressure cycle was applied in the FEA models to simulate a dent formed during construction. For unconstrained dent models, the indenter would be removed after indentation, and then a pressure cycle. The indentation step was applied in the FEA model by translating the indenter vertically downward.

Canadian standard, CSA Z662 (CSA, 2015), states that pipelines in Canada can operate not exceeding 80% of the specified minimum yield strength of the pipe (Equation (3.2)).

$$P = 80\% \times \frac{2St}{D} \tag{3.2}$$

where P is the maximum operating pressure (in MPa),

S is the specified minimum yield strength of the pipe (in MPa),

t is the wall thickness (in mm),

D is the outer diameter of the pipe (in mm).

CHAPTER 4: THREE-DIMENSIONAL STRAIN BASED MODEL FOR THE SEVERITY CHARACTERIZATION OF DENTED PIPELINES

4.1 Introduction

Dents are plastic deformations on a pipeline's cross-section caused by contact between the pipeline and digging equipment or rocks during installation, backfilling or following settlement over time. Such deformations can cause structural integrity or serviceability issues to the pipeline as they induce localized strains and stresses on the pipeline.

While some guidance is available in codes, regulations, and industry practices, dent assessment relies primarily on depth and conservative assumptions, which can lead to erroneous judgements on the performance of pipeline integrity programs.

Leading integrity techniques that entail accurate assessment of a wide range of materials, involving complex geometries under combined loading scenarios, rely on the use of numerical analysis methods such as finite element analysis (FEA) to evaluate the stress-strain conditions. However, FEA tends to be inefficient for the demand for computational time and resources for managing large pipeline systems due to prohibitively complex modeling and detailed analysis of each feature.

As such, there is substantial room for innovation and improvement to reduce the computation time for pipeline stress-strain analysis to support pipeline health management. Okoloekwe (2017) developed a robust mathematical approach to evaluate the strains in dented pipes by (1) using B-spline functions to interpolate the dented region of the pipe in a three-dimensional (3D) space and (2) implementing the dent strain equations on the interpolated

geometry**.** While the proposed technique provided a relatively good prediction of strain distribution benchmarked against the value by nonlinear FEA and the algorithm was simple to implement, the proposed technique consistently predicted conservative strain values compared to nonlinear FEA for all the models investigated.

The most common precedent for judging the severity of dents is the depth-based criterion, as adopted in many standards including the Canadian pipeline standards, CSA-Z662-16, which requires the repair of pipes with a plain dent deeper than 6% of the outside diameter (OD). Recent research has shown that the depth-based criterion for discerning the severity of a dent is indeed not sufficient as it might be unduly conservative in its predictions leading to unnecessary excavations (Gao et al., 2008). It is also possible for failures to occur in shallow dents as reported in the National Energy Board safety advisory, (Erikson, 2010). The American standards, ASME B31.8-2007 presents closed form expressions that can be used to evaluate the strains in a dented pipeline by discretizing the strain into components. A schematic representation of the strain components in a pipeline is shown in **Figure 12**.

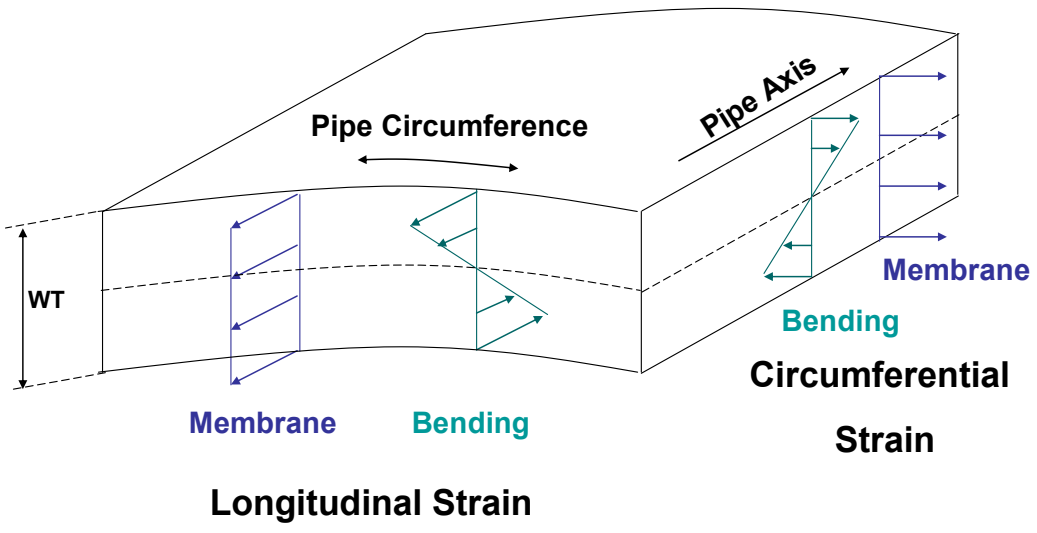


Figure 12 Strain components acting on a pipe wall (Lukasiewicz, et al 2006)

From **Figure 12**, it can be seen that the bending strains represented in Eq.(4.1) and (4.2) vary linearly along the thickness of the pipe wall, while the longitudinal membrane strains, represented in equation (3) which are as due to the extension of the pipeline in the longitudinal direction remains constant along the thickness of the pipe wall. The expressions for the circumferential bending strain (ε_1), the longitudinal bending strain (ε_2), and the longitudinal membrane strains (ε_3) at the peak of the dent as stipulated by ASME B 31.8-2016 standards are given in the Eqs. (4.1- 4.3)

$$\varepsilon_1 = \frac{t}{2}\left(\frac{1}{R_0} - \frac{1}{R_1}\right) \tag{4.1}$$

$$\varepsilon_2 = \frac{t}{2}\left(\frac{1}{R_2}\right) \tag{4.2}$$

$$\varepsilon_3 = \frac{1}{2}\left(\frac{d}{L}\right)^2 \tag{4.3}$$

where t is the pipe wall thickness, R_0 is the radius of the undeformed pipeline, d is the depth of the dent, L is the length of the dent, R_1 and R_2 are the external radii of curvature of the dent in the circumferential and longitudinal directions (see **Figure *13***), in the transverse and longitudinal planes through the dent. R_1 is positive when the dent partially flattens the pipe, in which case the curvature of the pipe surface in the transverse plane is in the same direction as the original surface radius of curvature R_0. Otherwise, if the dent is reentrant, R_1 is negative. The curvature R_2 as used in the code is generally always a negative value.

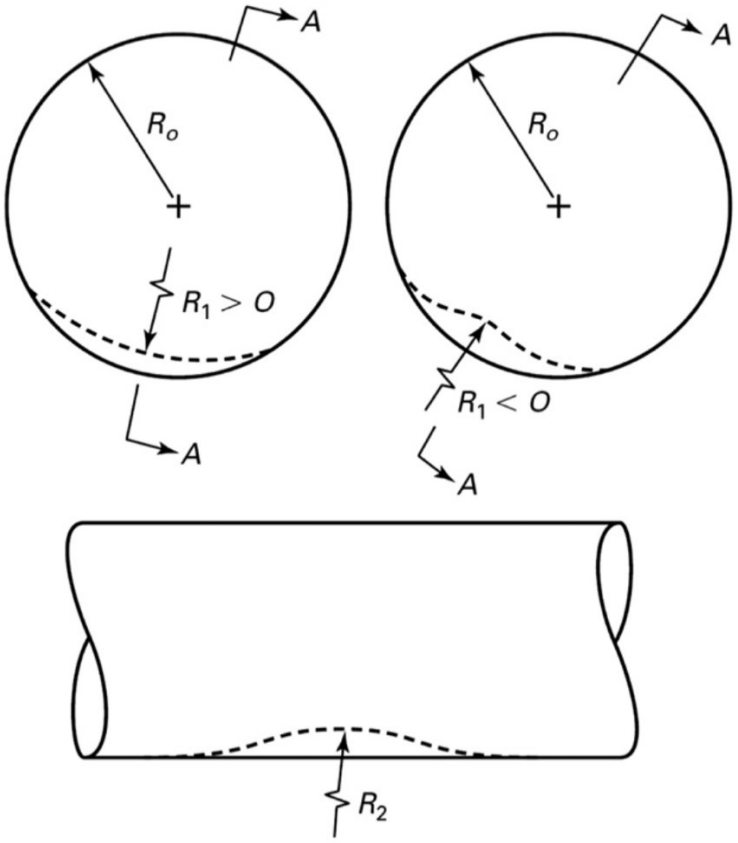


Figure 13 Dent geometry as defined in ASME B31.8 and reference in Noronha et al

These strain components are then combined accordingly to evaluate an equivalent total strain in the dented section as shown in Eq. (4.4) and (4.5).

$$\varepsilon_i = \sqrt{(\varepsilon_1)^2 - \varepsilon_1(\varepsilon_3 + \varepsilon_2) + (\varepsilon_3 + \varepsilon_2)^2} \tag{4.4}$$

$$\varepsilon_0 = \sqrt{(\varepsilon_1)^2 + \varepsilon_1(\varepsilon_3 - \varepsilon_2) + (\varepsilon_3 - \varepsilon_2)^2} \tag{4.5}$$

where ε_i and ε_0 are the strains in the inner and outer surfaces of the pipe wall respectively. The dent is considered as acceptable when the larger of the values ε_i and ε_0 is lower than the allowable strain limit. However, these equations were derived considering incorrect plane strain assumptions, and thus can lead to inaccurate results.

A more appropriate expression for the equivalent strain was presented by Noronha et al. (2010) considering the hypothesis that the strains in this region are mainly in the plastic range and the radial strains and the circumferential membrane strains are negligible. Hence, the expression for the equivalent strain state of the pipeline becomes Eq. (4.6)

$$\epsilon_{eqv} = \frac{2}{\sqrt{3}}\sqrt{\epsilon_I^2 + \epsilon_{II}^2 + \epsilon_I \epsilon_{II}} \tag{4.6}$$

where ϵ_I, ϵ_{II} and ϵ_{III} represent the principal strains in the longitudinal, circumferential and the radial directions.

Comparing the strains predicted by these models revealed that while the ASME B31.8 equations shown in Eq. (4.4 and 4.5), the expressions in Eq. (4.6) performed reasonably better than the previous equivalent strain expressions predicting strain values comparable to results from nonlinear FEA.

4.2 Methods

4.2.1 Modeling of Dents

FEA models for this study were developed using the commercial finite element software ABAQUS (Version 2016). The pipe segment is modeled as a 3D shell by extruding its cross section. The outer diameter (OD) of the pipe is 280 mm, the thickness of the wall is 8mm, and the length is 1100 mm. The indenter is modeled as a rigid analytical sphere. The parameters being investigated are 9 different dent depths (0%, 1%, 2%, 3%, 4%, 6%, 8%, 10%, and 12% OD) and 11 different diameters of the spherical indenter (25 mm, 30 mm…, 75 mm). As such a total of 99 numerical models were generated. Other boundary conditions used in the model include restrained rotation and translation at the ends of the pipe and restrained vertical translation at the bottom of the pipe. The contact surfaces in the models were defined with the master slave algorithm. The pipe was an X-60 pipe having a yield stress of 400 MPa. The pipe material was modeled as Elastic-

Plastic material having its elastic regime governed by a Young's Modulus of 200 GPa and a poisons ratio of 0.3. The pipe is an X-60 pipe having a yield stress, σ_y, of 400 MPa. The pipe was meshed with eight node brick elements.

The computation time of the numerical simulation was optimized by creating different levels of refinement, e.g., the region near to the dent was meshed with 5mm sized elements and then softened to 15 mm at 100 mm away from the dent's apex. Four elements were used through the wall thickness of the pipe.

The pipe deformation was modeled as restrained dents in unpressurised pipes. The indentation was applied by translating an indenter vertically downward, which created the pipe deformation.

4.2.2 Dent Profile Interpolation

According to the methodology proposed in the Appendix R of the ASME B31.8 Code, the calculation of the membrane strain in the longitudinal direction (ε_3) depends only on the length and depth of the dent. Therefore, the main question that arises is how to accurately determine the radii of curvature R_1 and R_2.

In this study, the radius of curvature of the dent profile is obtained as a mathematical function along the longitudinal and the circumferential directions of the deformed pipeline. This allows for evaluating the unique strain values at any point along the dent profile. The radii of curvature in the circumferential and the longitudinal direction are evaluated mathematically using the expressions shown in Eq.(4.7) and Eq.(4.8)

$$R_1 = \frac{\left(R^2+\left(\frac{dR}{d\theta}\right)^2\right)^{\frac{3}{2}}}{\left|R^2+2\left(\frac{dR}{d\theta}\right)^2-R\frac{d^2R}{d\theta^2}\right|} \tag{4.7}$$

$$R_2 = \frac{\left(1+\left(\frac{du_r}{dz}\right)^2\right)^{\frac{3}{2}}}{\left|\frac{d^2u_r}{dz^2}\right|} \tag{4.8}$$

where R is the inner radius of the pipe as reported by the inline inspection device, θ is the angular position of the pipe wall with interval $(-\pi \leq \theta \leq \pi)$ and Z is the longitudinal distance of the pipe with interval $(0 \leq Z \leq L)$, where L is the longitudinal distance from either side of the dent apex.

With the radius of curvatures obtained, the associated strains can then be easily evaluated using the closed form expressions defined in Eqs. (4.1)- (4.6). The computation of the strains employed in this study assumes that the tensile strains are developed at the internal surface of the pipe and the compressive strains at the external surface of the pipe.

The interpolation is done with the computing tool Mathematica which has inbuilt tools for the spline interpolation of surfaces. The data points that define the geometry of the dented surface of the pipeline are extracted from the numerical models and interpolated with B-spline curves as discussed in Noronha et al. (2005). The data points here refer to vector coordinates of the deformed pipeline which are extracted directly from the numerical model and converted into the cylindrical coordinate system as discussed in Luo and Chen (2000). The B-Spline curves are polynomials between a pair of data points with components defined in such a way that some level of smoothness up to a particular derivative is attained.

The coordinates are extracted at 64 points along the circumference of the pipeline at every 10 mm interval. The spline functions used for the surface interpolation are equipped with second order continuity so as to generate a differentiable mathematical surface for the dent.

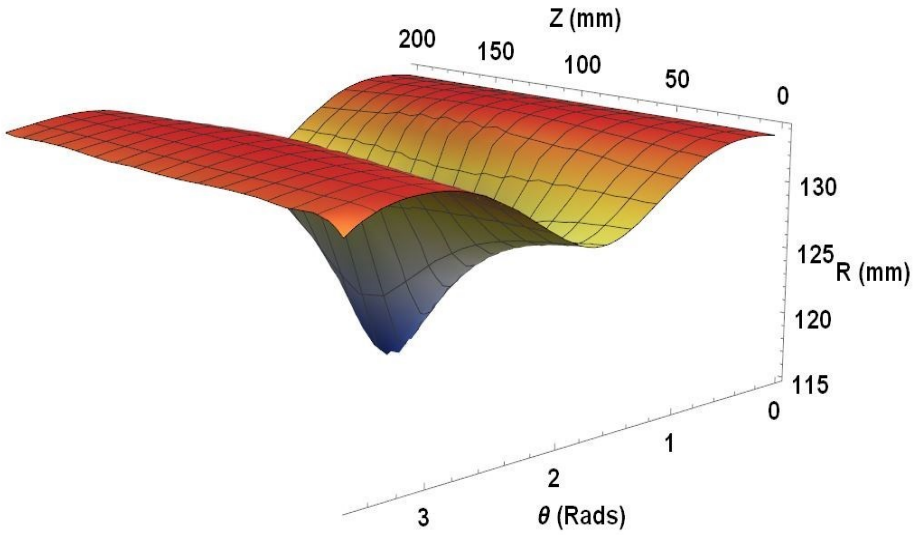
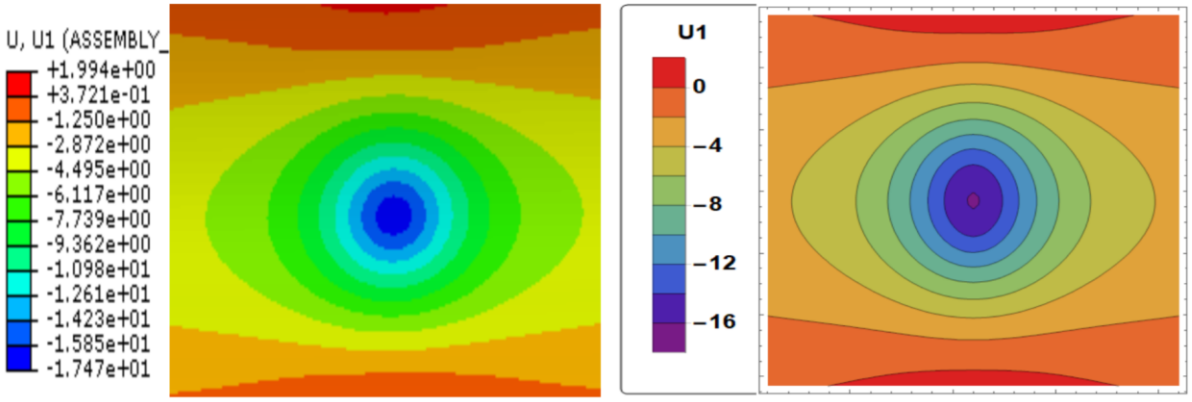


Figure 14 Interpolated dent surface of the SD6 model

4.3 Results

4.4.1 Deformation Analysis

The deformation analysis is performed by solving the expressions for the directional displacements along the entire region of the dented pipe. From **Figure 15**, the contours represent the radial displacement distribution by both the numerical and the analytical models at the internal surface of a 6% OD spherical dent-depth pipeline. The analytically evaluated radial displacement, (U1) of the SD6 model is obtained using the expressions shown in Eq. (4.12) and the radial displacement from the numerical model is obtained from the FEA simulation and the results are shown in the **Figure 15**.



(a) **(b)**

Figure 15 SD6-radial displacement contours: (a) numerical model and (b) analytical model

A good correlation in the radial displacement predicted by the analytical and the numerical methods is observed with both techniques predicting a peak radial displacement of 17.5 mm, respectively.

The circumferential displacement (U2) and the longitudinal displacement (U3) distribution predicted by both numerical and analytical models are also be compared and shows a good correlation.

4.4.2 Sensor Investigation

The dent surface is generated by extracting the radial displacements and the angular positions directly from the nodes of the numerical model. The resolution of the data points was selected to interpolate the dent surface as to simulate deformation obtained from an ILI tool. For the previously analyzed models, the data coordinates used to define the geometry of the deformed pipe were extracted at 64 points along the circumference of the pipe from the finite element displacement at every 10mm interval in longitudinal direction. The number of sensors along the circumference of the pipeline was varied to account for the different accuracies obtainable in inline inspection tools. The sensor numbers investigated were64, 32, 16 and 8 sensors along the

circumference of the pipeline. The equivalent plastic strains were evaluated using the equations proposed in Lukasiewicz et al., 2006). The strains compared are the longitudinal strains (LE33) and the equivalent plastic strains (PEEQ).

The distribution of equivalent plastic strains along the pipe transverse section near the dent is presented in **Figure 16**. For the B-Spline methodology, each different strain curves was obtained considering the interpolation of the dent geometry with an decreasing number of known points, corresponding, respectively, to a deformation ILI tool with 64, 32, 16 and 8 sensors. These results are compared with the FE results. And the **Figure 17** plots of the maximum equivalent plastic strain values between the FEA and the analytical model developed with 64, 32, 16, and 8 sensors of the R25mm indenter model.

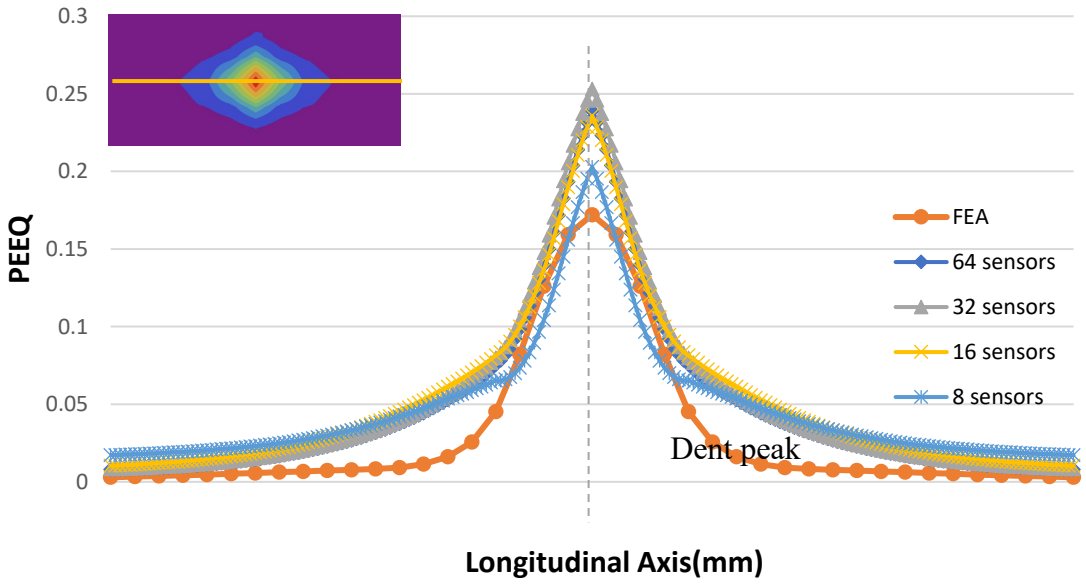


Figure 16 Equivalent plastic strains developed at (a) numerical model (b) 64-sensors, (c) 32-sensors, (d) 16-sensors, and (e) 8-sensor

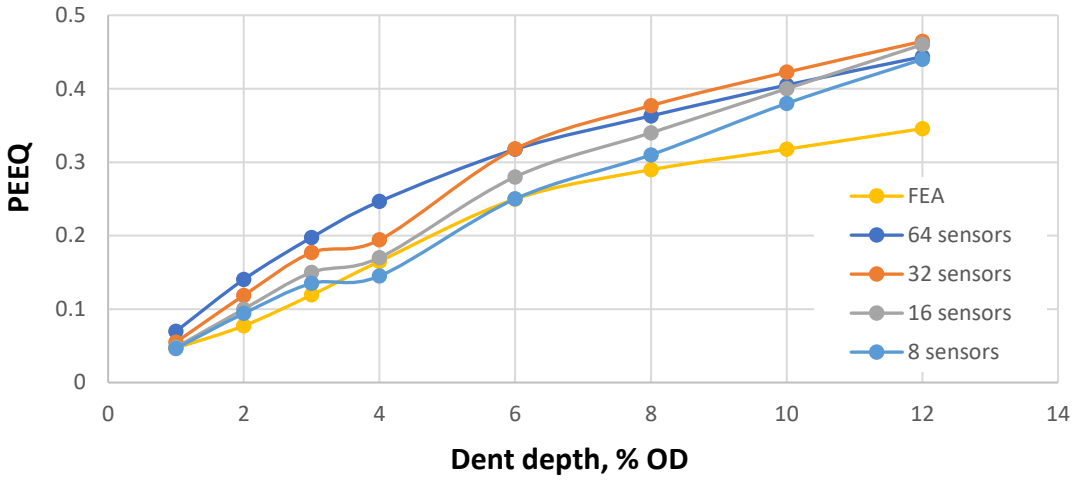


Figure 17 The maximum equivalent plastic strain values of the FEA model and the analytical models developed with 64, 32, 16, and 8 sensors

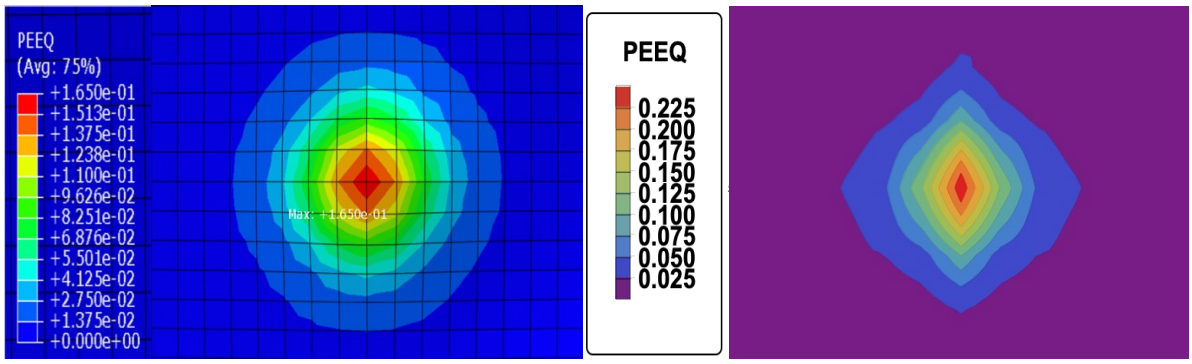
It can be deduced that the greater number of data points used for the interpolation, the more conservative strains were predicted. Again, the accuracy of the strain results increases with the number of sensors; however, while the strain curves along the pipe transverse section is well represented by only 16 sensors, it can be seen that a higher number of sensors is needed to provide a more conservative estimation for the maximum PEEQ at the dent apex. And considering the attainability by inline inspection devices, the 32 sensors along the circumference of the pipe method was just chosen for the following analysis.

4.4.3 Strain Analysis

The contours presented in this section (**Figure 18**) are inner surface views from the dent depth of 6% OD 50mm-diameter indenter case

A good correlation in both displacements and strains predicted by both the numerical and the analytical models is observed.

The strain distribution predicted using the mathematical model proposed is benchmarked against the strains predicted by nonlinear FEA. The contour plots of the equivalent plastic strains (PEEQ) generated by nonlinear FEA and the mathematical technique.



(a) (b)

Figure 18 The equivalent plastic strains (PEEQ) of (a) numerical model and (b) analytical model

The parametric analysis is performed in order to investigate the correlation between the strain values and the dent depth and dent geometry.

The equivalent strains referred to in this study have been termed as “plastic” as the contribution of the elastic component of the deformation has been ignored. This was also done to ease the comparison of the analytical models and the numerical models.

Table 2 The maximum equivalent plastic strain values of the FEA model and the analytical models developed with 32 sensors

R(mm) D/t		1%	2%	3%	4%	6%	8%	10%	12%
25	FEA	0.047	0.0773	0.1193	0.1650	0.2498	0.2897	0.3177	0.3456
	Analytical	0.055	0.1186	0.1770	0.1942	0.3180	0.3770	0.4225	0.4645
	Difference	0.008	0.0413	0.0577	0.0292	0.0682	0.0873	0.1048	0.1189
30	FEA	0.03942	0.0766	0.1144	0.1617	0.2259	0.2593	0.2802	0.2903
	Analytical	0.0509	0.1036	0.1692	0.2253	0.2935	0.3540	0.3950	0.4250
	Difference	0.01148	0.02704	0.0548	0.06355	0.0676	0.0947	0.1148	0.1347
35	FEA	0.03527	0.0760	0.1106	0.1564	0.2058	0.2386	0.2469	0.2502
	Analytical	0.0481	0.1136	0.1638	0.2165	0.2758	0.3320	0.3660	0.3858
	Difference	0.01283	0.03756	0.0532	0.0601	0.06995	0.0934	0.1191	0.13555
40	FEA	0.03203	0.0754	0.1059	0.1472	0.1936	0.2185	0.2191	0.2198
	Analytical	0.0464	0.1102	0.1596	0.2038	0.2634	0.3110	0.3306	0.3190
	Difference	0.01437	0.0348	0.0537	0.0566	0.0698	0.0925	0.1115	0.0992
45	FEA	0.03003	0.0732	0.1042	0.1398	0.1831	0.1968	0.1967	0.1974
	Analytical	0.045	0.1062	0.1570	0.1942	0.2530	0.2866	0.3000	0.3100
	Difference	0.01497	0.03303	0.0528	0.0544	0.0699	0.0898	0.1033	0.1126
50	FEA	0.02809	0.0716	0.1009	0.1327	0.1721	0.1790	0.1793	0.1795
	Analytical	0.441	0.1036	0.1522	0.1942	0.2430	0.2670	0.2750	0.2900
	Difference	0.41291	0.03201	0.0513	0.0615	0.0709	0.088	0.0957	0.1105
55	FEA	0.02765	0.0704	0.0971	0.1268	0.1607	0.1668	0.1647	0.1629
	Analytical	0.0434	0.1014	0.1470	0.1792	0.2310	0.2525	0.2550	0.2650
	Difference	0.01575	0.03098	0.04986	0.0524	0.0703	0.0857	0.0903	0.1021
60	FEA	0.02735	0.0696	0.0941	0.1207	0.1504	0.1550	0.1518	0.1487
	Analytical	0.0427	0.0998	0.1428	0.1734	0.2198	0.2376	0.2409	0.2510
	Difference	0.01535	0.03018	0.04874	0.0527	0.0694	0.0826	0.0891	0.1023
65	FEA	0.02703	0.0688	0.0898	0.1166	0.1418	0.1442	0.1417	0.1382
	Analytical	0.0423	0.0984	0.1382	0.1700	0.2109	0.2228	0.2250	0.2350
	Difference	0.01527	0.02959	0.04838	0.0534	0.0691	0.0786	0.0833	0.0968
70	FEA	0.02677	0.0681	0.0864	0.1130	0.1344	0.1361	0.1324	0.1301
	Analytical	0.0422	0.0971	0.1343	0.1676	0.2025	0.2118	0.2149	0.2200
	Difference	0.01543	0.02898	0.04795	0.0546	0.0681	0.0757	0.082525	0.0899
75	FEA	0.02654	0.0675	0.0832	0.1087	0.1289	0.1286	0.1243	0.1216
	Analytical	0.0418	0.0964	0.1309	0.1636	0.1958	0.2026	0.2057	0.2169
	Difference	0.01526	0.02888	0.04767	0.0549	0.0669	0.074	0.0814	0.0953

From the chart generated, it is observed that the proposed technique consistently predicts conservative strain magnitudes and fairly similar strain distributions compared to nonlinear FEA for all the models investigated.

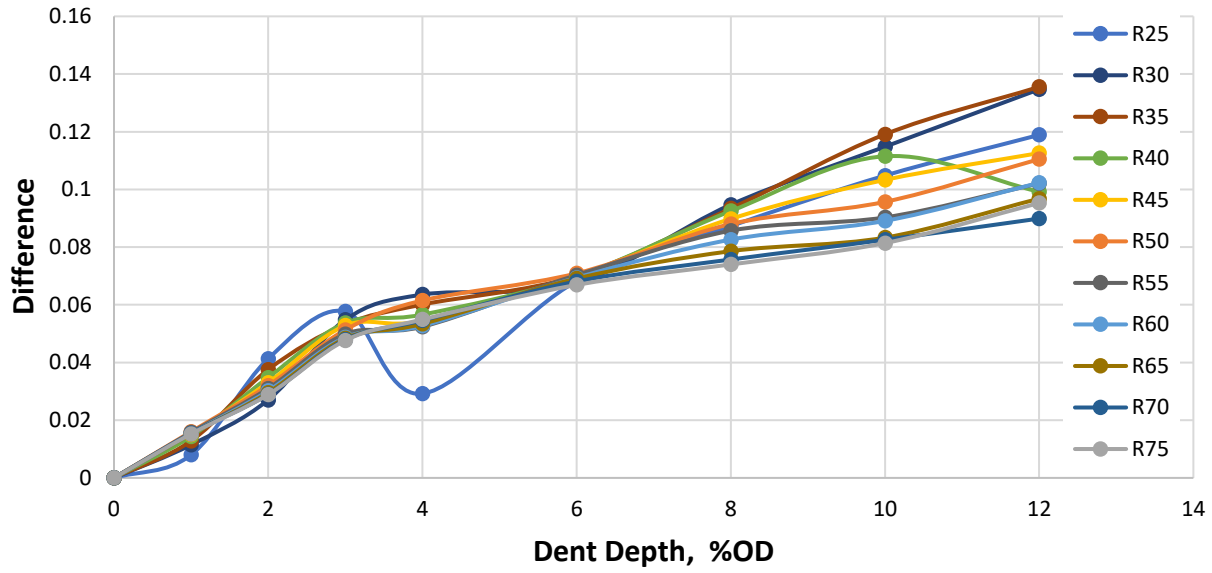


Figure 19 Plots of the difference between the FEA model and the analytical model in terms of the maximum PEEQ

It will be interesting to investigate how these closed-form expressions perform when used for the analysis of unrestrained dents subjected to internal pressure cycles.

Figure 19 shows the value of the maximum PEEQ of the analytical model minus the nonlinear FEA model, where the modeled dents are restrained and the pipes are not pressurized, thus excluding the fatigue cycling and the rebounding of the dented surface, and the analytical model is observed to be more conservative than the nonlinear FEA irrespective of the dent depth. The level of conservatism in the strain analytical model is also seen to have some correlation with the dent depth as the magnitude increases with increasing depths.

CHAPTER 5: GAUSSIAN PROCESS REGRESSION-BASED ACCURACY ASSESSMENT OF ANALYTICAL MODEL FOR DENT ANALYSIS

As demonstrated in Chapter 4, the strain behavior of dented pipe can vary significantly due to changes in dent depth and indenter size. However, the process of matching the displacement profile obtained from an FEA model to interpolated with B-spline functions can be a time-consuming, iterative process.

The objective of Chapter 5 is to demonstrate the feasibility of using Gaussian Process Regression to efficiently and accurately estimate the strain error of the analytical model without running the FEA for those untried dent configurations.

5.1 Definition

The GP model is a probabilistic, non-parametric model, which differs from most of the other parametric models by fitting the parameters of the selected basis functions. By contrast, GP searches for the relationship from the observed data (Petelin et al., 2013).

A Gaussian Process is a collection of random variables, any finite number of which have (consistent) joint Gaussian distributions.

A Gaussian process is fully specified by its mean function $m(x)$ and covariance function $k(x, x')$. This is a natural generalization of the Gaussian distribution whose mean and covariance is a vector and matrix, respectively. The Gaussian distribution is over vectors, whereas the Gaussian process is over functions:

$$f(x) \sim GP\big(m(x), k(x, x')\big) \tag{5.1}$$

meaning: "the function $f(x)$ is distributed as a GP with mean function $m(x)$ and covariance function $k(x, x')$".

5.2 Posterior Gaussian Process

The GP approach finds a distribution over the possible functions $f(x)$ that are consistent with the observed data. As with all Bayesian methods, it begins with a prior distribution and updates this as data points are observed, producing the posterior distribution over functions.

We first consider a simple 1-d regression problem, mapping from an input x (dent depth) to an output $f(x)$ (the error of the max PEEQ between analytical and FEA model). In **Figure 20(a)**, ten sample functions were drawn at random from the prior distribution over functions specified by a particular Gaussian process assuming that the average value over the sample functions at each x is zero. This prior is taken to represent our prior beliefs over the kinds of functions, before seeing any data.

In **Figure 20 (b)**, after given a dataset of three observations $D = \{(x_1, f(x_1)), (x_2, f(x_2)), (x_3, f(x_3))\}$, and only consider functions exactly pass though these three data points. The solid lines show sample functions which are consistent with D, and the dashed line depicts the mean value of such functions. Notice the uncertainty is reduced close to the observations because the combination of the prior and the dataset leads to the posterior distribution over functions.

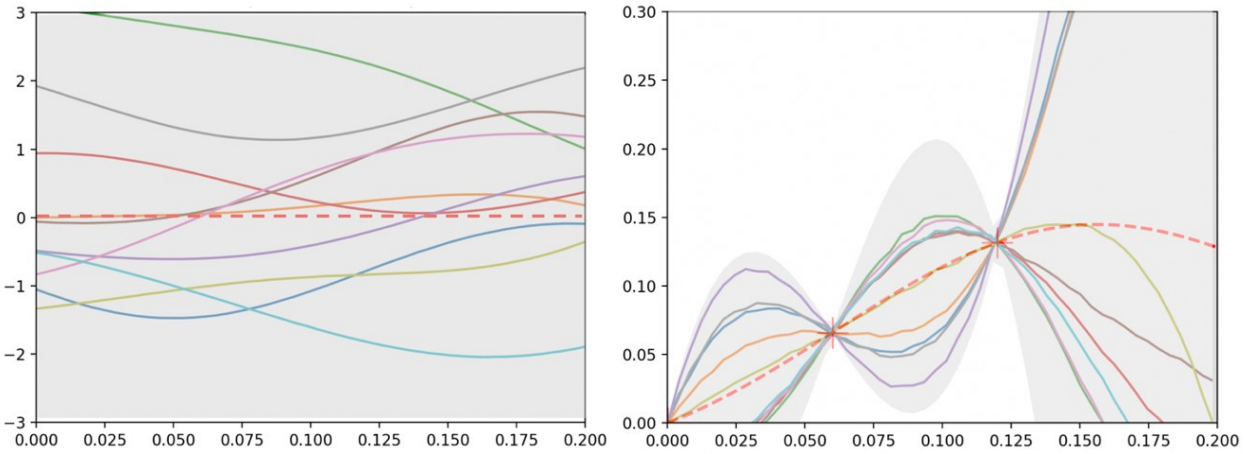


Figure 20 Gaussian process inference

Figure 20 (a) shows ten samples drawn from the prior distribution. Panel (b) shows the situation after three data points have been observed. The mean prediction is shown as the dashed line and ten samples from the posterior are shown as solid lines. In both plots the shaded region denotes twice the standard deviation at each input value x.

In the above example, how to define distributions over functions using GP is shown. Now, one of the primary goals computing the posterior is that it can be used to make predictions for unknown test cases. Initially, we will consider the simple special case where the observations are noise free, that is we know $\{(x_i, f_i)|i = 1, \ldots, n\}$. According to the Gaussian Process properties, the observed (denoted collectively by $\boldsymbol{f}$) and unobserved (denoted collectively by f_*) function values follow a joint Gaussian distribution:

$$\begin{pmatrix} f \\ f_* \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \mu_* \end{pmatrix}, \begin{pmatrix} K & K_* \\ K_*^T & K_{**} \end{pmatrix}\right) \tag{5.2}$$

Here, $\mu = m(x_i)$ is the training means and analogously for the test means μ_*. $K = K(x, x)$ is the $n \times n_*$ covariance matrix to indicate the similarity of the observed x in the training data set observed (e.g., the data summarized in **Table 2**). $K_* = K(x, x_*)$ characterizes the similarity of the

observed x in the training data set to the test data point whose output values are to be estimated. $K_{**} = K(x_*, x_*)$ represents the similarity of the test data points in the test data set. The objective is to derive the posterior probability density function $p(f_*|x_*, x, f)$ given f and we are assuming that f and f_* together are jointly Gaussian as defined above.

A series of matrix algebra (the formula for conditioning a joint Gaussian distribution):

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left(\begin{pmatrix} a \\ b \end{pmatrix}, \begin{pmatrix} A & C \\ C^T & B \end{pmatrix}\right) \Longrightarrow x|y \sim N(a + CB^{-1}(y - b),\ A - CB^{-1}C^T) \tag{5.3}$$

can get us from the joint distribution $p(f, f_*)$ to the conditional $p(f_*|f)$. We are of course interested in the conditional probability $p(f_*|f)$: "given the data, how likely is a certain prediction for f_*?", which follows a Gaussian distribution:

$$f_*|f \sim N(\mu_* + K_*^T K^{-1}(f - \mu),\ K_{**} - K_*^T K^{-1} K_*) \tag{5.4}$$

Function values f_* (corresponding to test inputs x_*) can be sampled from the joint posterior distribution by evaluating the mean and covariance matrix from Eq. (5.4) and generating samples.

5.4 Prediction using Noisy Observations

Note that it is typical for more realistic modelling situations that we do not have access to function values themselves, but only noisy versions. The most common assumption is that of additive Gaussian noise in the outputs and such noise is easily taken into account. Assuming the additive independent identically distributed Gaussian noise $\varepsilon \sim N(0, \sigma_n^2)$, each observation y can be thought of as related to an underlying function $f(x)$ through a Gaussian noise model: $y = f(x) + \varepsilon$. The prior on the noisy observations becomes

$$cov(y_p, y_q) = k(x_p, x_q) + \sigma_n^2 \delta_{pq}$$

where δ_{pq} is a Kronecker delta which is one if $p = q$ and zero otherwise. It follows from the independence assumption about the noise for different observations, that a diagonal matrix is added, in comparison to the noise free case. Introducing the noise term in Eq. (5.2) we can write the joint distribution of the observed target values y and the function values at the test locations f_* under the prior. Specifically, if the prior mean function $\mu = 0$, we have that:

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim N\left(0, \begin{pmatrix} K+\sigma_n^2 I & K_* \\ K_*^T & K_{**} \end{pmatrix}\right) \tag{5.5}$$

Deriving the conditional distribution corresponding to Eq. (5.4) we arrive at the key predictive equations for Gaussian process regression.

So, the posterior on f_* is:

$$f_*|y \sim N(K_*^T K_y^{-1} y, K_{**} - K_*^T K_y^{-1} K_*) \tag{5.6}$$

where $K_y = K + \sigma_n^2 I$ is the covariance matrix for the noisy targets y rather than the underlying function f.

Note first that the mean prediction is a linear combination of observations y; this is sometimes referred to as a linear predictor.

Note also that the variance does not depend on the observed targets, but only on the inputs; this is a property of the Gaussian distribution. The first term K_{**} is simply the prior covariance; from that is subtracted a (positive) term, representing the information the observations gives us about the function. We can very simply compute the predictive distribution of test targets y_* by adding $\sigma_n^2 I$ to the variance in the expression for $cov(f_*)$.

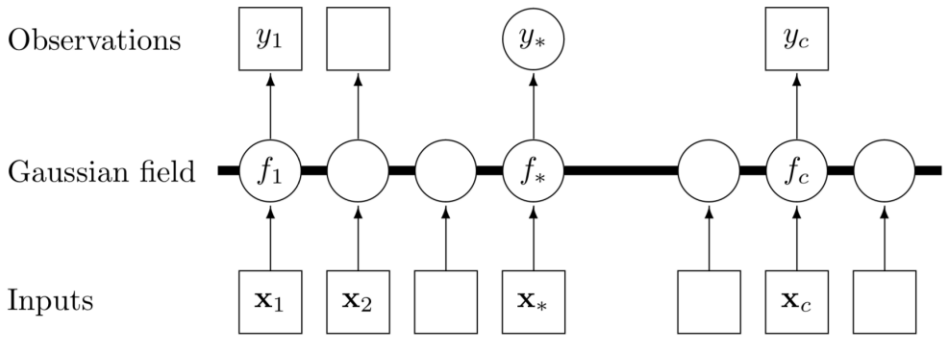


Figure 21 Graphical model for a GP for regression (Rasmussen & Williams, 2006)

A graphical model representation of a GP is given in **Figure 21**, the chain graph shows of GPR. Squares represent observed variables and circles represent unknowns. The thick horizontal bar represents a set of fully connected nodes. Note that an observation y_i is conditionally independent of all other nodes given the corresponding latent variable, f_i. Because of the marginalization property of GPs addition of further inputs, x, latent variables, f, and unobserved targets, y_*, does not change the distribution of any other variables.

5.5 Covariance Function

We have seen that a covariance function is the crucial ingredient in a Gaussian process predictor, as it the key factor that controls the properties of a Gaussian process and encodes our assumptions about the function which we wish to learn. From a slightly different viewpoint, the covariance matrix can smooth the function we are interested in by ensuring that values that are close in the input space will produce output values that are close, and thus training points that are near to a test point should be informative about the prediction at that point. A key fact of Gaussian processes is that they can be completely defined by their second-order statistics (Bishop, 2006). Thus, if a Gaussian process is assumed to have mean zero, defining the covariance function completely defines the process' behaviour. Basic aspects that can be defined through the

covariance function are the process' stationarity, isotropy, smoothness and periodicity (Barber, 2012)

Stationarity refers to the process' behaviour regarding the separation of any two points x and x'. If the process is stationary, it depends on their separation, $x - x'$. Thus, it is invariant to translations in the input space. While if non-stationary it depends on the actual position of the points x and x'. For example, the squared exponential covariance function is stationary.

If the process depends only on $|x - x'|$, the Euclidean distance (not the direction) between x and x', then the process is considered isotropic. It is thus invariant to all rigid motions. A process that is concurrently stationary and isotropic is considered to be homogeneous; in practice these properties reflect the differences (or rather the lack of them) in the behaviour of the process given the location of the observer.

If a covariance function depends only on x and x' through $x \cdot x'$ we call it a dot product covariance function. Dot product covariance functions are invariant to a rotation of the coordinates about the origin, but not translations.

Ultimately Gaussian processes translate as taking priors on functions and the smoothness of these priors can be induced by the covariance function (Barber, 2012). It is a basic assumption that points with inputs x which are close are likely to have similar target values y also, then the concept of continuity is present. If we wish to allow for significant displacement then we might choose a rougher covariance function. Extreme examples of the behaviour are the Ornstein–Uhlenbeck covariance function and the squared exponential where the former is never differentiable and the latter infinitely differentiable.

Periodicity refers to inducing periodic patterns within the behaviour of the process. Formally, this is achieved by mapping one-dimensional input variable x to a two-dimensional vector $u(x) = (cos(x), sin(x))$ (MacKay, 1998).

5.5.1 Squared Exponential Covariance Function

The squared exponential (SE), or Gaussian covariance function is the most commonly used kernel, as shown below,

$$k_{SE}(x, x') = \sigma_f^2 e^{-\frac{(x-x')^2}{2l^2}} \tag{5.7}$$

Note it has two hyperparameters: the length scale l and signal covariance σ_f. The length scale l determines the strength of correlation between points. It briefly defines how far apart the input values x_i can be for the response values to become uncorrelated. And σ_f specifies deviation away from the mean function. Both parameters control the smoothness of the functions estimated by a GP.

Although the squared exponential is probably the most widely used kernel within the kernel machines field, its strong smoothness assumptions are unrealistic for modelling many physical processes because the covariance function is infinitely differentiable, which means that the GP with this covariance function has mean square derivatives of all orders.

Stein [1999] argues that such strong smoothness assumptions are unrealistic for modelling many physical processes and recommends the Matérn class (see below).

The Matérn class of covariance functions is given by

$$k_{Matern}(r) = \frac{2^{1-v}}{\Gamma(v)l^{2v}}\left(\frac{\sqrt{2v}r}{l}\right)^{v} K_v\left(\frac{\sqrt{2v}r}{l}\right) \tag{5.8}$$

Table 3 lists several commonly used covariance functions. The covariances are written either as a function of x and x', or as a function of $r = |x - x'|$. 'S' and 'ND' indicate whether the covariance functions are stationary and nondegenerate.

Table 3 Summary of several commonly used covariance functions (Rasmussen & Williams, 2006)

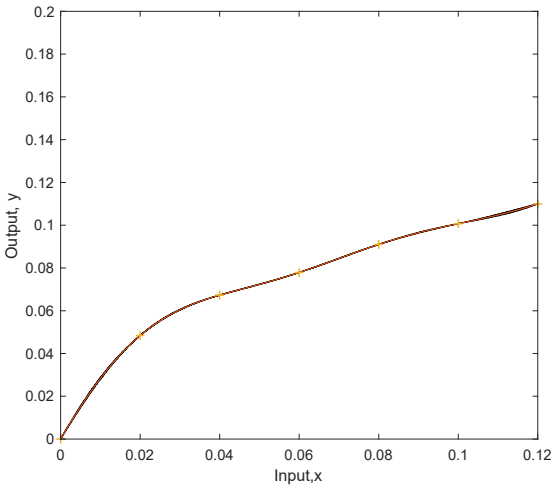
covariance function	expression	S	ND
constant	σ_0^2	√	
linear	$\sum_{d=1}^{D} \sigma_d^2 x_d x'_d$		
polynomial	$(\mathbf{x} \cdot \mathbf{x}' + \sigma_0^2)^p$		
squared exponential	$\exp(-\frac{r^2}{2\ell^2})$	√	√
Matérn	$\frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\ell} r\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}}{\ell} r\right)$	√	√
exponential	$\exp(-\frac{r}{\ell})$	√	√
γ-exponential	$\exp\left(-(\frac{r}{\ell})^{\gamma}\right)$	√	√
rational quadratic	$(1 + \frac{r^2}{2\alpha\ell^2})^{-\alpha}$	√	√
neural network	$\sin^{-1}\left(\frac{2\tilde{\mathbf{x}}^\top \Sigma \tilde{\mathbf{x}}'}{\sqrt{(1+2\tilde{\mathbf{x}}^\top \Sigma \tilde{\mathbf{x}})(1+2\tilde{\mathbf{x}}'^\top \Sigma \tilde{\mathbf{x}}')}}\right)$		√

5.6 Varying the Hyperparameters

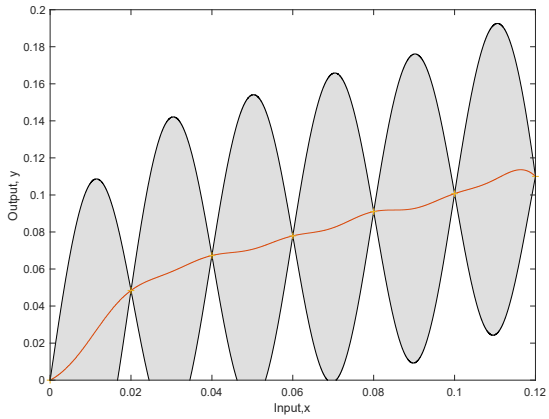
Typically, the covariance functions that we use will have some free parameters. For example, the squared-exponential covariance function in one dimension has the following form

$$k_y = k_f + \sigma_n^2 \delta(x, x') = \sigma_f^2 e^{-\frac{(x-x')^2}{2l^2}} + \sigma_n^2 \delta(x, x') \tag{5.9}$$

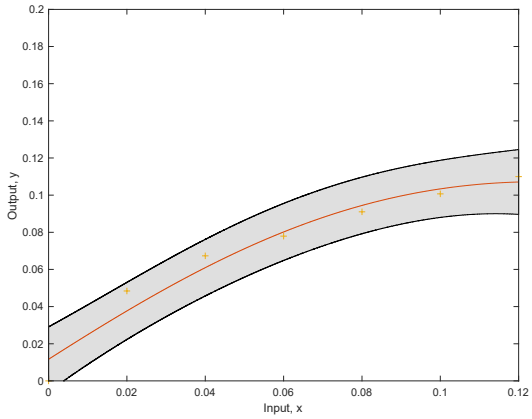
The covariance is denoted k_y as it is for the noisy targets y rather than for the underlying function f. The reliability of our regression is dependent on how well we select the covariance function and its hyperparameters θ $(l, \sigma_f^2, \sigma_n^2)$.



(a) $\theta = (0.045, 0.079, 0.000003)$



(b) $\theta = (0.01, 0.072, 0)$



(c) $\theta = (0.135, 0.091, 0.0067)$

Figure 22 Varying the hyperparameters on GP prediction

Figure 22 shows three functions drawn from posterior, given 7 training points (as shown by the + symbol)). 10000 data points are generated from a GP with three different sets of hyperparameters θ (l, σ_f, σ_n), as shown by red line. The shaded grey area represents the 95% confidence region for the underlying function f.

In next section we will consider the problem of learning in Gaussian processes by finding suitable hyperparameters for the covariance function. However, in this section our aim is more

simply to explore the effects of varying the hyperparameters on GP prediction. Here are the three plots from the one-dimensional training input study which only fixed the radius of indenter to 45mm while varying the dent depth from 0 to 12% of the OD. Consider the data shown by + signs in **Figure 22**. Here the training inputs x are 7 evenly spaced points between 7 different dent depths (0%, 2%, 4%, 6%, 8%, 10%, and 12% OD) and outputs y are the corresponding error of PEEQ between the mathematical model and the analytical model. The figure also shows the 2 standard-deviation error bars for the predictions obtained using these values of the hyperparameters, as per Eq. (5.6). Notice how the error bars would reduce close to the observations.

The function in **Figure 22(b)** varies too rapidly as we make predictions with a process with l = 0.01 and the remaining two parameters were set by optimizing the marginal likelihood, as explained in next section. In this case the noise parameter is reduced 0, so the underlying function f have to sharply varies to exactly pass through all the given data points. Notice that the error bars in **Figure 22(b)** grow rapidly away from the datapoints because of the short length-scale.

Rather, we can set the length-scale longer, for example to l = 0.135 then we get the result shown in 0**Figure 22(c)**. In this case, the noise parameter set by optimizing the marginal likelihood the data increased to σ_n = 0.0067 and the prediction mean function varies slowly with a lot of noise, which can be observed at the two datapoints x = 0.02 and x = 0.03.

In conclusion, the function in **Figure 22(b)** has little noise but too wiggly. In contrast, the function in **Figure 22(c)** is pretty smooth while has too much noise. In this case the marginal likelihood gives a clear preference for the hyperparameters set in **Figure 22(a)** over the other two alternatives.

5.7 Bayesian Model Comparison

Bayes' rule is given by:

$$posterior \propto likelihood \times prior$$

$$p(w|y,x) \propto p(y|x,w) \times p(w)$$

It is common to use a hierarchical specification of models. At the lowest level are the parameters, w. For example, the parameters could be the parameters in a linear model, or the weights in a neural network model.

The Bayesian view of model comparison involves the use of probabilities to represent uncertainty in the choice of the model. We would like to compare a (discrete) set of L models $\{M_i\}$, where $i = 1,2,\dots,L$

We specify the prior distribution over the different models $p(M_i)$, the posterior over the parameters is given by Bayes' rule

$$\text{posterior}=\frac{likelihood \times prior}{marginal\ likelihood} \qquad p(M_i|y,x) = \frac{p(y|x,M_i)p(M_i)}{p(y|x)} \tag{5.9}$$

where the normalizing constant in the denominator of Eq. (5.9) $p(y|x)$ is independent of the model structures, and called the marginal likelihood (or model evidence), and is given by

$$p(y|x) = \int p(y|x,M_i)p(M_i)dM_i \tag{5.10}$$

At the second level are hyperparameters θ which control the distribution of the parameters at the bottom level. For example, the "weight decay" term in a neural network, or the "ridge" term in ridge regression are hyperparameters.

At the next level, we analogously express the posterior over the hyperparameters, where the likelihood from the model level plays the role of the marginal likelihood

$$p(\theta|y,x,M_i) = \frac{p(y|x,\theta,M_i)p(\theta|M_i)}{p(y|x,M_i)} \tag{5.11}$$

where $p(\theta|M_i)$ is the hyper-prior (the prior for the hyperparameters). The normalizing constant is given by

$$p(y|x, M_i) = \int p(y|x, \theta, M_i)p(\theta|M_i)d\theta \tag{5.12}$$

At the bottom level, the posterior over the parameters is given by Bayes' rule

$$p(w|y, x, \theta, M_i) = \frac{p(y|x,w,M_i)p(w|\theta,M_i)}{p(y|x,\theta,M_i)} \tag{5.13}$$

where $p(y|x, w, M_i)$is the likelihood and $p(w|\theta, M_i)$ is the parameter prior. The prior encodes as a probability distribution our knowledge about the parameters prior to seeing the data. If we have only vague prior information about the parameters, then the prior distribution is chosen to be broad to reflect this. The posterior combines the information from the prior and the data (through the likelihood). The normalizing constant in the denominator of Eq. (5.13) $p(y|x, \theta, M_i)$ is independent of the parameters, and called the marginal likelihood (or evidence), and is given by

$$p(y|x, \theta, M_i) = \int p(y|x, w, M_i)p(w|\theta, M_i)dw \tag{5.14}$$

We note that the implementation of Bayesian inference calls for the evaluation of several integrals. In practice, especially the evaluation of the integral in Eq. (5.12) may be difficult. The fully Bayesian predictive distribution is then given by marginalizing over model parameters as well as hyperparameters:

$$p(y|x, M_i) = \int p(y|x, \theta, M_i)p(\theta|M_i)d\theta = \iint p(y|x, w, M_i)p(w|\theta, M_i)p(\theta|M_i)dwd\theta \tag{5.15}$$

However, this integral is intractable (even when everything is Gaussian). Need to approximate.

Note: the fully Bayesian approach is to integrate over the posterior distribution for $\{\theta, w\}$. This can be done by Markov chain Monte Carlo (MCMC) methods. For now, we will use evidence approximation, which is much faster.

If we assume that the posterior over hyperparameters θ is sharply picked, we can approximate:

$$p(y|x, M_i) \approx p(y|x, \theta, M_i) = \int p(y|x, w, M_i)p(w|\theta, M_i)dw \tag{5.16}$$

So, we integrate out parameters and maximize the marginal likelihood in w.r.t. the hyperparameters, θ. This approximation is known as type II maximum likelihood (ML-II), empirical Bayes or Evidence Approximation. Of course, one should be careful with such an optimization step, since it opens up the possibility of overfitting, especially if there are many hyperparameters. The integral in Eq. (5.12) can then be approximated using a local expansion around the maximum (the Laplace approximation). The prior over models M_i in Eq. (5.9) is often taken to be flat, so that a priori we do not favour one model over another.

$$p(M_i|y, x) \propto p(y|x, M_i) \tag{5.17}$$

In this case, the probability for the model is proportional to the expression from Eq. (5.12).

$$p(M_i|y, x) \propto p(y|x, \theta, M_i)p(\theta|M_i) \tag{5.18}$$

The marginal likelihood from Eq. (5.14) involving the integral over the parameter space and there is no weighting parameter which needs to be set by some external method such as cross validation. This is a feature of the marginal likelihood that it automatically incorporates a trade-of between model fit and model complexity. This is the reason why the marginal likelihood is valuable in solving the model selection problem. **Figure 23** illustrates how the automatic trade-off comes about.

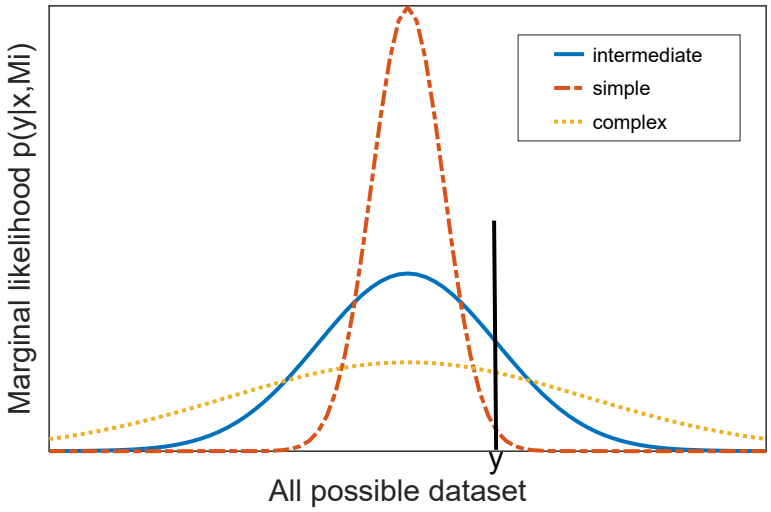


Figure 23 Occam's razor is automatic (Rasmussen and Williams, 2006)

On the horizontal axis is an abstract representation of all possible datasets (of a particular size). The vertical axis, the marginal likelihood $p(y|x, M_i)$ represents the probability of the data given the model. Three different complex models are shown. For a particular dataset indicated by y, the model with intermediate complexity has the largest marginal likelihood.

In **Figure 23** we show a schematic of the behaviour of the marginal likelihood for three different model complexities. Let the number of data points n and the inputs x be fixed. For a particular dataset indicated by y, the marginal likelihood prefers a model of intermediate complexity over too simple or too complex alternatives. The simple model cannot fit the data well, whereas the more complex model spreads its predictive probability and so assigns relatively small probability to any one of them. A more complex model can account for many more data sets than a simple model, but since the probabilities have to integrate to unity, this means more complex models are automatically penalized more. The figure illustrates why the marginal likelihood doesn't simply favour the models that fit the training data the best.

5.8 Marginal Likelihood

The best way to determine the hyperparameters is to learn them from the data. We would like to maximize the likelihood of a prediction given the training data and the parameters.

In the previous section we saw how to update the prior Gaussian process in the light of training data. This is useful if we have enough prior information about a dataset at hand to confidently specify prior mean and covariance functions. However, the availability of such detailed prior information is not the typical case in machine learning applications. In order for the GP techniques to be of value in practice, in this section our aim is to explore how such parameters can be inferred or learned from the data, based on either Bayesian methods (by optimizing the marginal likelihood).

we must be able to choose between different mean and covariance functions in the light of the data. This process will be referred to as training the GP model.

In the light of typically vague prior information, we use a hierarchical prior, where the mean and covariance functions are parameterized in terms of hyperparameters. For example, we could use a generalization of Eq. (5.1):

$$f(x) \sim GP\big(m(x), k(x,x')\big),$$

$$m(x) = ax^2 + bx + c, \text{and } k(x,x') = \sigma_f^2 e^{-\frac{(x-x')^2}{2l^2}} + \sigma_n^2 \delta(x,x') \tag{5.19}$$

where we have introduced hyperparameters $\theta = \{a, b, c, \sigma_f, \sigma_n, l\}$. The purpose of this hierarchical specification is that it allows us to specify vague prior information in a simple way. For example, we've stated assuming the function to be close to a second order polynomial. In fact, the discrepancy between the polynomial and the data is a smooth function plus independent Gaussian noise, but again we're don't need exactly to specify the characteristic length scale l or

the magnitudes of the two contributions. We want to be able to make inferences about all of the hyperparameters in the light of the data.

In the previous section we saw the reliability of our regression is dependent on how well we select the covariance function. Our maximum posteriori estimate of θ occurs when $p(\theta|x,y)$ is at its greatest. Bayes' theorem tells us that, assuming you have little prior knowledge about what θ should be, this corresponds to maximizing $log\ p(y|x,\theta)$. Fortunately, this is not difficult, since by assumption the distribution of the data is Gaussian. And we now explicitly write the marginal likelihood conditioned on the hyperparameters θ.

$$L = \log p(y|x,\theta) = -\frac{1}{2}\log|K_y| - \frac{1}{2}(y-\mu)^T|K_y|^{-1}(y-\mu) - \frac{n}{2}\log(2\pi) \tag{5.20}$$

We call Eq. (5.20) the log marginal likelihood since it is the likelihood of a non-parametric model and it is obtained through marginalization over the latent function.

We can now find the values of the hyperparameters which optimizes the marginal likelihood based on its partial derivatives which are easily evaluated:

$$\frac{\partial L}{\partial \theta_m} = -(y-\mu)^T\Sigma^{-1}\frac{\partial m}{\partial \theta_m}$$

$$\frac{\partial L}{\partial \theta_k} = \frac{1}{2}trace\left(K_y{}^{-1}\frac{\partial K_y}{\partial \theta_k}\right) + \frac{1}{2}(y-\mu)^T\frac{\partial K_y}{\partial \theta_k}K_y{}^{-1}\frac{\partial K_y}{\partial \theta_k}(y-\mu) \tag{5.21}$$

where θ_m and θ_k are used to indicate hyperparameters of the mean and covariance functions respectively. Eq. (5.21) can conveniently be used in conjunction with a numerical optimization routine such as conjugate gradients to find good hyperparameter settings.

Due to the fact that the Gaussian process is a non-parametric model, the marginal likelihood behaves somewhat differently to what one might expect from experience with parametric models. Note first, that it is in fact very easy for the model to fit the training data exactly: simply set the noise level $\sigma_n{}^2$ to zero, and the model produce a mean predictive function which agrees exactly

with the training points. However, this is not the typical behavior when optimizing the marginal likelihood.

Indeed, the log marginal likelihood from Eq. (5.20) consists of three terms: The first term, $-\frac{1}{2}\log|K_y|$, a complexity penalty term measures and penalizes the complexity of the model depending only on the covariance function. The second term $-\frac{1}{2}(y-\mu)^T|K_y|^{-1}(y-\mu)$, plays the role of a data-fit. It is the only term involving the observed targets y. The third term $\frac{n}{2}\log(2\pi)$ is a log normalization constant, independent of the data.

We've seen in this section how we, via a hierarchical specification of the prior, can express prior knowledge in a convenient way, and how we can learn values of hyperparameters via optimization of the marginal likelihood. This can be done using some gradient based optimization. Also, we've seen how the marginal likelihood automatically incorporates Occam's razor; this property of great practical importance, since it simplifies training a lot.

5.8 Training a Gaussian Process using GPML package

The GPML toolbox can be obtained from http://gaussianprocess.org/gpml/code/matlab/doc/index.html is an Octave 3.2.x and MATLAB 7.x implementation of inference and prediction in Gaussian process (GP) models. It implements algorithms discussed in Rasmussen & Williams: *Gaussian Processes for Machine Learning*, the MIT press, 2006

Gaussian process conducted in this research borrows heavily from "The GPML Toolbox version 4.2". Using the GPML package is simple, there is only one single function to call: gp, it does posterior inference, learns hyperparameters, computes the marginal likelihood and makes predictions. Generally, the gp function takes the following arguments: a hyperparameter struct, an

inference method, a mean function, a covariance function, a likelihood function, training inputs, training targets, and possibly test cases.

Here, the error in predicting the maximum PEEQ from the analytical model compared with FEA is considered as the output function $f(x)$ and x indicates the dent depth and the diameter of the spherical indenter.

A GP modelling tool, GPML, is used to learn the error function of interest in this study. **Figure 24** illustrates GP modelling is used to quantify the prediction error in the analytical model here. The training data set consists of the inputs x at 99 grid points, i.e., 9 different dent depths (0%, 1%, 2%, 3%, 4%, 6%, 8%, 10%, and 12% OD) and 11 different diameters of the spherical indenter (25mm, 30mm, …,75mm) and outputs y as the corresponding difference in the maximum PEEQ predicted between the FE model and the analytical model.

From **Figure 24**, 99 dots indicate training data. The dotted surface is an estimation of the error of the predicted PEEQ of analytical models for different dent depth and geometry. Pointwise 95% confidence intervals are shaded in blue and green. We can basically get any error value from any dent depth and indenter radius combination in this domain.

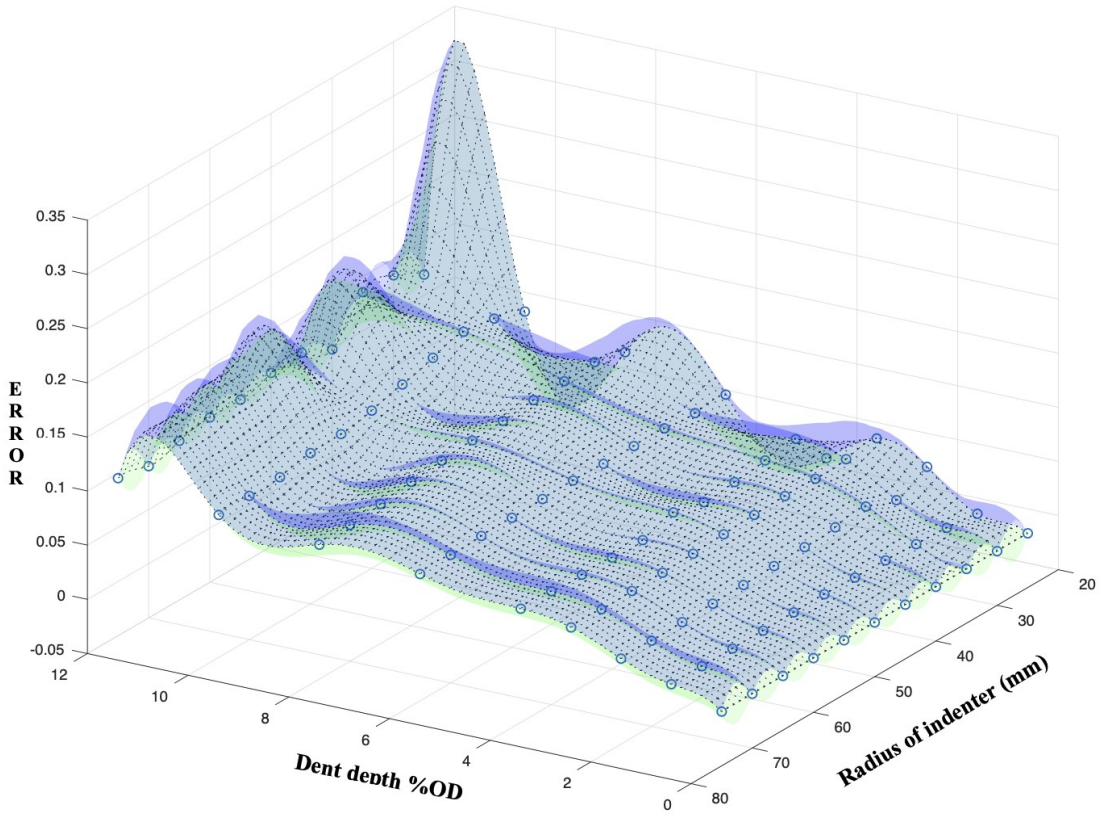


Figure 24 Two- dimensional GP model

Two- dimensional GP model for modeling the error of the predicted PEEQ of analytical models with varying dent depth and geometry

CHAPTER 6: DEVELOPMENT OF RANDOM FOREST

This section gives a general overview of random forests and some related concepts about the method. And a proposal for building an RF model with the database we already have.

6.1 Introduction

Random forest is an algorithm that integrates multiple trees through the idea of Ensemble Learning. Its basic unit is decision tree, while its essence belongs to a large branch of machine Learning, Ensemble Learning method.

In fact, from an intuitive point of view, each decision tree is a classifier (assuming that it is now aimed at classification problems), so for an input sample, n trees will have n classification results. The random forest integrates all the classification voting results and specifies the category with the most votes as the final output, which is the simplest idea of Bagging.

6.2 Algorithm

6.2.1 Decision Tree Learning

A decision tree is a tree- structure model in which each internal node represents a test on an attribute, each branch represents a test output, and each leaf represents a category. Decision trees where the target variable can take continuous values are called regression trees. Common decision tree algorithms include C4.5 (successor of ID3), ID3 (Iterative Dichotomiser 3) and CART (Classification and Regression Tree).

Decision tree learning is a predictive method commonly used in statistics, data mining and machine learning. It uses a decision tree as a predictive model to predict the value of a target variable based on several input variables.

Some techniques, often called ensemble methods, construct multiple decision trees:

6.2.2 Ensemble Learning

Ensemble learning solves a particular computational intelligence problem by generating and combination of multiple models. In each case, a final decision is made by combining the individual decisions of several classifiers/models. These predictions are combined into a single forecast and are therefore superior to any single forecast.

A random forest is a subclass of ensemble learning, which integrates many decision trees into forest and uses them together to predict the final result.

6.2.3 Bootstrap Aggregating

The decision tree works well especially if they are of small depth. However, a decision tree can have great depth in practice are more prone to overfitting to learn highly irregular patterns resulting in a larger variance. This decision trees' shortcoming of overfitting to their training set is corrected by the Random Forest model. In order to reducing the variance, the original training data is randomly sampled-with-replacement generating small subsets of data (see the image below). These subsets are also known as bootstrap samples. These bootstrap samples are then fed as training data to many decision trees of large depths. Each of these decision trees is trained separately on these bootstrap samples. This aggregation of decision trees is called the Random Forest ensemble.

The concluding result of the ensemble model is determined by counting a majority vote from all the decision trees. This concept is known as Bagging or Bootstrap Aggregation. Since each uncorrelated decision tree takes a different set of training data as input, the sensitivity to noise of a single tree predictions in its training set do not impact the final result obtained from the aggregation of decision trees. Therefore, bagging as a concept reduces variance without increasing the bias of the complete ensemble.

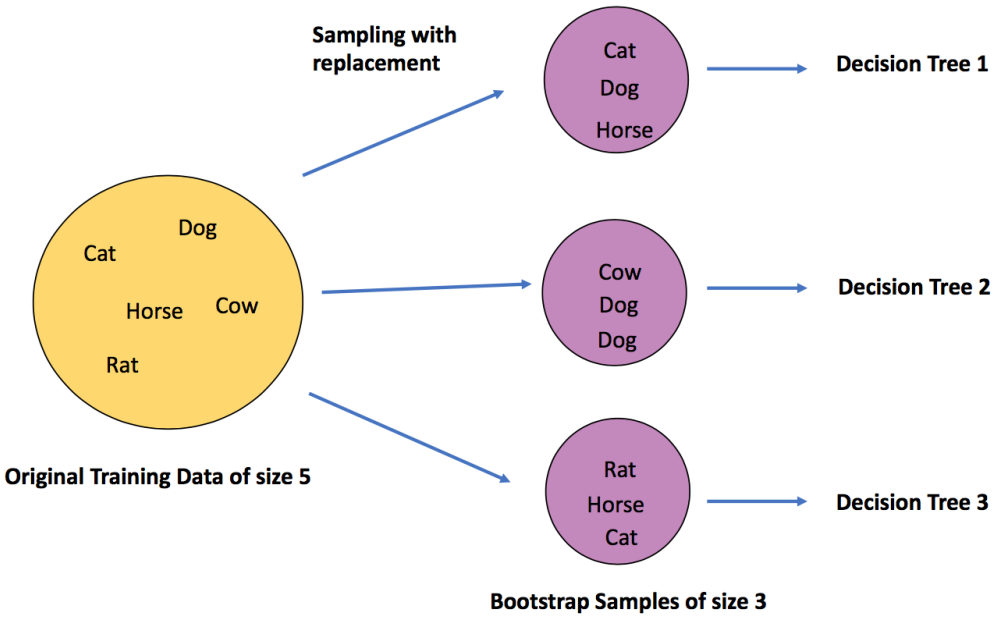


Figure 25 Generation of bootstrap samples with replacement

(Source: https://miro.medium.com/max/1400/1*ixvrbH45K8CcNZaj98JGuA.png)

The training algorithm for random forests applies the general technique of bagging to tree learners. Given a training set $D = ((x_1, y_1) \ldots, (x_n, y_n))$ of size n, bagging M times to generate new training sets D_i of size n', by sampling from original training set randomly and with replacement.

By sampling with replacement, some observations may be repeated in each D_i. If $n' = n$, then for large n, the probability of not picking a sample in a random draw is $\lim_{n\to\infty} \left(1 - \frac{1}{n}\right)^n = e^{-1} = 0.368$. Therefore, about 36.8 % of total training data are available as OOB sample for each decision tree and hence it can be used for evaluating or validating the random forest model. This kind of sample is known as a bootstrap sample. Then, m models are fitted using the above m bootstrap samples and combined by averaging the output (for regression) or voting (for classification). After training, predictions for unseen samples x' can be made by averaging the predictions (for regression):

$$\hat{f}(x') = \frac{1}{M}\sum_{m=1}^{M} f_m(x') \tag{6.1}$$

or by taking the majority voting (for classification).

Additionally, an estimate of the uncertainty of the prediction can be made as the standard deviation of the predictions from all the individual regression trees on x':

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - \hat{f}(x'))^2}{B-1}} \tag{6.2}$$

The number of samples, m, is a free parameter. Typically, a few hundred to several thousand trees are used, depending on the size and nature of the training set. An optimal number of trees m can be found using cross-validation, or by observing the out-of-bag error:

6.2.4 Out of Bag Error

Out-of-bag (OOB) error, also called out-of-bag estimate, is a method of measuring the prediction error of random forests and other machine learning models utilizing bootstrap aggregating (bagging) to sub-sample data samples used for training. OOB is the mean prediction error on each training sample x_i, calculated using predictions from the trees that do not contain x_i in their bootstrap sample.

6.2.5 Advantages of Random Forests

- The accuracy rate is excellent among the current algorithms
- Able to run efficiently on large data sets with numerous variables
- It can process input samples with high dimensional characteristics without dimensionality reduction
- Be able to estimate the variable importance

- In the generation process, an unbiased estimate of the internal generation error can be obtained
- It offers a superior method for working with missing data

6.3 Train Ensemble of Bagged Regression Trees

Statistics and Machine Learning Toolbox™ provides objects of ensemble learning algorithms, including bagging, random space, and various boosting algorithms.

Three objects Bootstrap Aggregation (Bagging) and Random Forest:

- ClassificationBaggedEnsemble created by fitcensemble for classification
- RegressionBaggedEnsemble created by fitrensemble for regression
- TreeBagger created by TreeBagger for classification and regression

Bootstrap aggregation (bagging) is a type of ensemble learning. To bag a weak learner such as a decision tree on a data set, generate many bootstrap replicas of the data set and grow decision trees on the replicas. Obtain each bootstrap replica by randomly selecting n out of n observations with replacement, where n is the data set size. In addition, every tree in the ensemble can randomly select predictors for each decision split, a technique called random forest known to improve the accuracy of bagged trees. By default, the number of predictors to select at random for each split is equal to the square root of the number of predictors for classification, and 1/3 of the number of predictors for regression. After training a model, you can find the predicted response of a trained ensemble for new data by using the predict function. predict takes an average over predictions from individual trees.

By default, the minimum number of observations per leaf for bagged trees is set to 1 for classification and 5 for regression. Trees grown with the default leaf size are usually very deep.

These settings are close to optimal for the predictive power of an ensemble. Often you can grow trees with larger leaves without losing predictive power. Doing so reduces training and prediction time, as well as memory usage for the trained ensemble. You can control the minimum number of observations per leaf by using the 'MinLeafSize'.

Several features of bagged decision trees make them a unique algorithm. Drawing n out of n observations with replacement omits about 36.8% of observations for each decision tree. These omitted observations are called “out-of-bag” observations. TreeBagger has properties and object functions, whose names start with oob, that use out-of-bag observations.

- oobPredict estimates the out-of-bag prediction by averaging predictions from all trees in the ensemble for which the observation is out of bag.
- oobError computes the misclassification probability (for classification trees) or mean squared error (for regression trees) for out-of-bag observations in the training data
- OOBPermutedPredictorDeltaError estimates of feature importance by permutation of out-of-bag predictor observations for random forest of classification trees. The software randomly permutes out-of-bag data across one variable at a time and estimates the increase in the out-of-bag error due to this permutation. The larger the increase, the more important the feature.

6.4 Application to Lager Dataset

The following example shows the workflow for regression using TreeBagger.

In order to improve the understanding of the factors that affect the prediction’s accuracy. The process described in Section 4.2 was repeated in this section but with a larger training dataset where other 5 variables, Operating pressure, outside diameter, Wall thickness, pipe grade and

indentation conditions were also modified to enrich the database to train the random forest. The properties that were adjusted is shown in **Table 4**.

Table 4 List of different pipe and indenter properties used in FEA and analytical model

Operating pressure (MPa)	0
	4
Outside diameter (mm)	260
	280
	300
Wall thickness (mm)	6
	7
	8
Pipe grade	X52
	X60
Indentation conditions	Constrained
	Unconstrained
Indentation depth (% of OD)	2, 3, 4, …,12
Indenter radius (mm)	25, 30, 35, …, 75

It is also worth mentioning that from the renewed database, the analytical model shows non-conservative prediction when taking operating pressure (4 MPa) into consideration and the results are shown in **Figure 26**, where the $\overline{\varepsilon^p}$ from the analytical method is plotted on the y-axis, while the $\overline{\varepsilon^p}$ from FEA is plotted on the x-axis, with the blue line through the center of the plot marking the locus of equal equivalent plastic strain between the analytical model and FEA. At lower strains, the predictions of the equivalent plastic strain from the analytical model match the FEA, but a deviation is observed at higher strains where the analytical model becomes less

conservative predicting strains that are lower than those recorded by the FEA, which is in agreement with what other researchers have found (Woo, 2019). Okoloekwe (2018) attributed the fact that the discrepancies in the magnitude of the strain were due to the limited degrees of freedom of the analytical model.

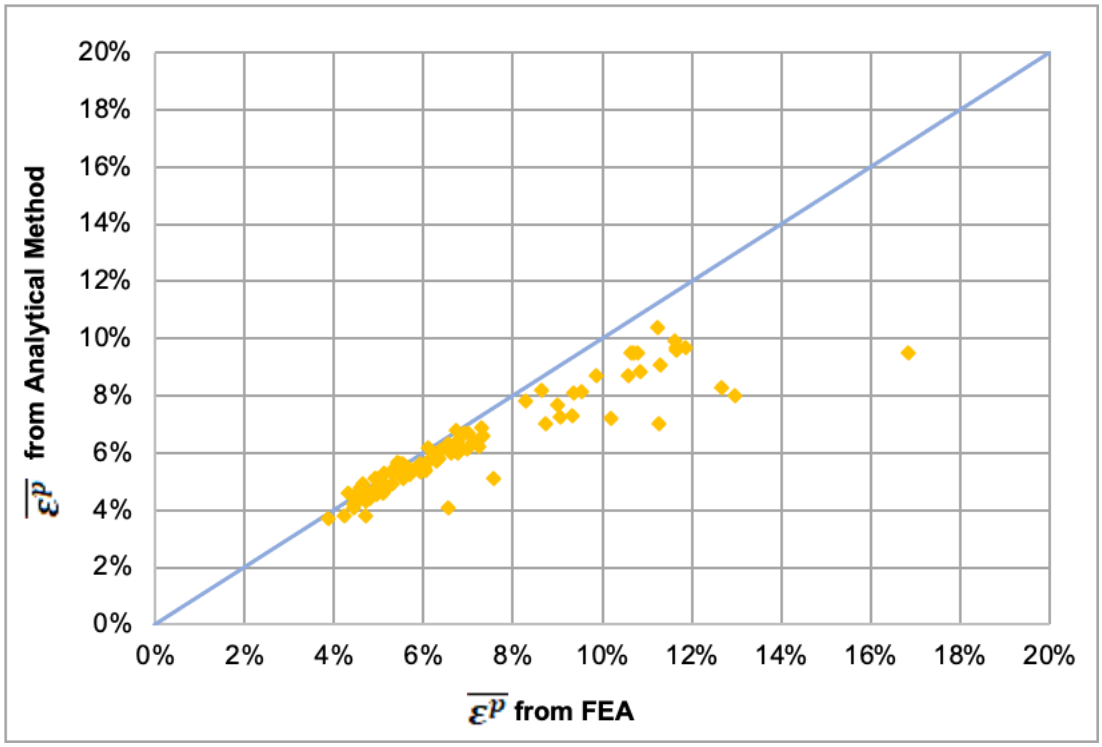


Figure 26 Comparison of maximum PEEQ predicted from the analytical model vs FEA with operating pressure

Use the TreeBagger algorithm on matlab. Train a random forest of 100 regression trees using the database of the strain error with 152 observations, 7 predictors, and 1 response which is 'PEEQ'. Note, the 'Pipe grade' and 'Indentation conditions' are categorical variables while others are numerical variables.

Load the data set and split it into predictor and response arrays

```
load difference
features = X(:,(1:7));
classLables = X(:8);
```

Grow number of ‘nTrees’ trees. For regression, it is best to set the minimal leaf size to 5 and select one third of the total number of features for each decision split at random. These settings are defaults for TreeBagger used for regression.

```
b = TreeBagger(nTrees,features,classLables,'Method','R', ...
'OOBPredictorImportance','On',...
   'CategoricalPredictors',X(:7),'MinLeafSize',5);
```

Plot the increase in MSE due to permuting out-of-bag observations across each input variable. The OOBPermutedPredictorDeltaError array stores the increase in MSE averaged over all trees in the ensemble and divided by the standard deviation taken over the trees, for each variable. The larger this value, the more important the variable.

```
figure
bar(b.OOBPermutedPredictorDeltaError)
xlabel('Feature Number')
ylabel('Out-of-Bag Feature Importance')
```

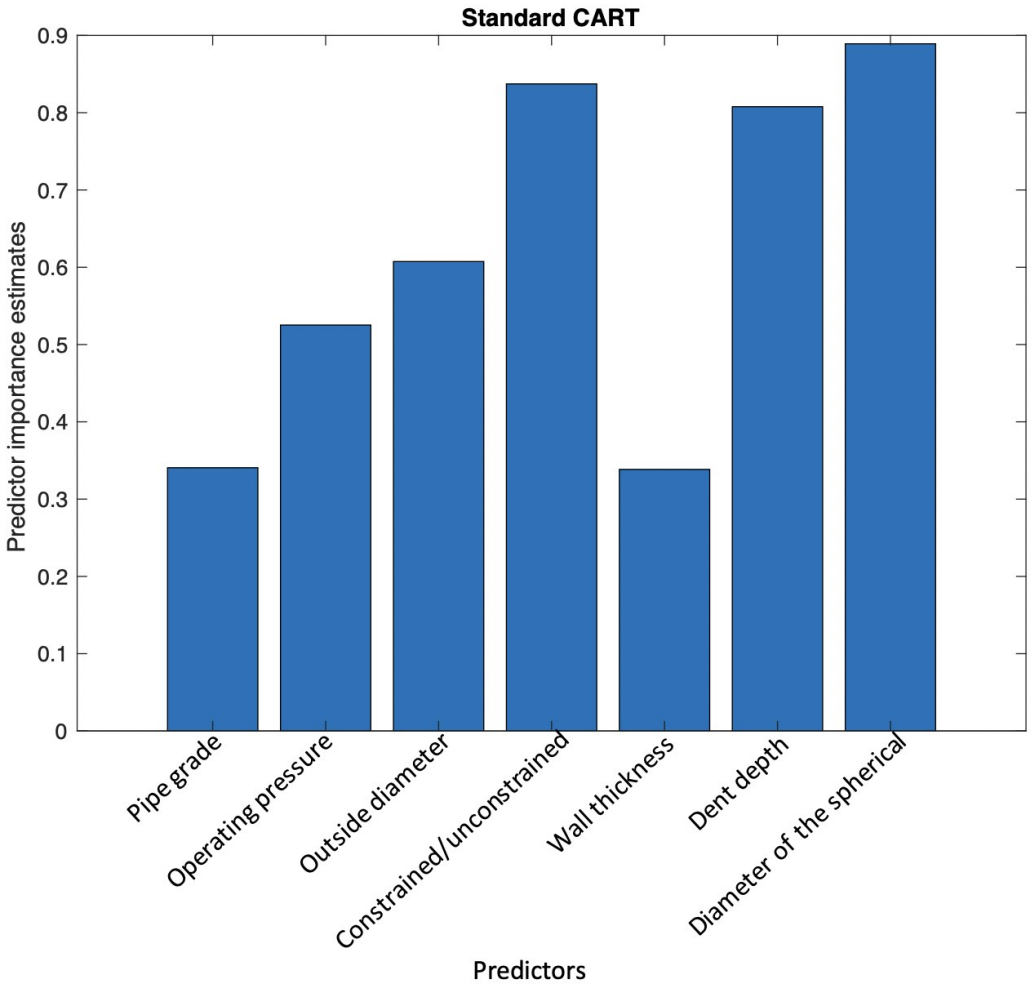


Figure 27 Estimating Feature Importance

From Figure 27, in this case, the most important feature is the indenter diameter followed closely by indentation condition.

CHAPTER 7: CONCLUSIONS AND FUTURE RESEARCH

7.1 Summary and Conclusions

While dents are a prevalent concern across liquid pipeline systems, current regulations are simply based on depth and interaction with other features. Accurate and efficient methodologies for assessing dents is not widely agreed upon in industry. The benefit of having an effective integrity assessment method for dents would allow the prevention of pipeline failure, which can negatively affect the environment and the health and safety of people. While some researchers have proposed the use of finite element analysis to assess dents, this method requires significant computational time and is impractical for system-wide applications.

In this thesis, a technique for analytically evaluating the strains in dented pipelines, based on the coordinates of the geometric profile of the dent (Okoloekwe, 2017) was conducted. And the strains predicted from the said method are benchmarked against the strains predicted from a numerical model generated using nonlinear finite element analysis (FEA). The good agreement obtained in the strains predicted by the developed model and FEA indicates a possibility of conducting in-depth strain analysis of thin- walled structures. In order to estimate the uncertainty in the analytical model, a series of nonlinear FEA pipe indentation simulations were conducted using the finite element analysis tool, ABAQUS and compared with the analytical prediction.

Lastly, Gaussian Process Regression (GPR)-based machine learning method was used to reduce uncertainty and quantifying the error in terms of the maximum equivalent plastic strain (PEEQ) benchmarked against the values by nonlinear FEA. These results show that GPR can be used to predict the maximum strains and stresses in dented regions with similar accuracy to what is achieved using FEA, which has already been proven to be an accurate representation of real-life by other researchers. The use of machine learning methods creates the added benefit of achieving

results in far less time than using FEA alone and more accurate than analytical model could be feasible for system-wide application from a time and resource perspective. Thus, another potential method using random forest was proposed as well.

7.2 Recommendations for Future Research

While this research demonstrated that it is feasible to assess dents with a combination of FEA and machine learning methods, the research did not investigate all the different pipe properties and dent types that can be found on a pipeline system. For training and validating the model, many parameters were held constant in this study. Future work could investigate the adjustment of several different parameters at once (for example, the GPR would cover a range of pipe diameters, wall thicknesses, and grades in addition to various indenter sizes), and evaluate if the prediction model could still produce accurate results. This would allow for practical applications to pipeline systems that have many lines with various properties.

Furthermore, this research only explored plain, symmetric dents with single apexes. These complications need to be investigated and resolved. Dent interaction with other features such as other dents, metal loss, and/or cracks has the potential to increase a dent's severity than if the dent was not interacting with any other features. In addition, dents can have irregular shapes: for example, they can have multiple apexes, or their peaks may not align with the longitudinal axis of the pipe. These features in the FEA models as well as with the GPR needs to be investigated further.

In this research, a strain hardening exponent of 12 was assumed in the Ramberg-Osgood equation to approximate a stress-strain curve for all FEA models. The effect of this assumption could be assessed against the use of stress-strain curves from experimental coupon tests.

Future research could further improve the accuracy and efficiency of the proposed methods. The GPR was trained using the GPML package in the commercially available software, MATLAB.

More accurate results could potentially be achieved by using more advanced training algorithms to train the GPR. In order to improve efficiency further, cloud computing could be implemented with the proposed automation techniques to allow the FEA models to run simultaneously and build the GPR training database faster.

This research was simply focused on finding the deterministic maximum values for strains and stresses, the research could be expanded to utilize reliability techniques and account for uncertainties in the measurement of the profile as well as other properties such as the pipe's material properties.

Another important step would be to compare the results obtained with the other approaches: random forest or even other machine learning methods like support vector regression (SVR). It would be very important to use the same data set with both approaches and make a direct comparison between them.

REFERENCES

Adeeb, S.,2011. Introduction to Solid Mechanics and Finite Element Analysis Using Mathematica. Kendall Hunt.

Alexander, C.R., 1999, Analysis of Dented Pipeline Considering Constrained and Unconstrained Dent Configurations," Proceedings of the Energy Sources Technology Conference and Exhibition, Houston, Texas, USA.

ASME B31.8 Gas Transmission and Distribution Piping Systems. 2003. ASME International, New York, NY.

ASME B31.8 Gas Transmission and Distribution Piping Systems. 2007. ASME International, New York, NY.

ASME B31.8 Gas Transmission and Distribution Piping Systems. 2016. ASME International, New York, NY.

Baker, M., 2004. Integrity Management Program–Dent Study. Department of Transportation, Office of Pipeline Safety, TTO Number 10.

Barber, David., 2012. Bayesian Reasoning and Machine Learning. Cambridge University Press.

Barbian, A. and Beller, M., 2012, In-line Inspection of High Pressure Transmission Pipelines: State-of-the-Art and Future Trends. Proceedings of the 18th World Conference on Nondestructive Testing, Durban, South Africa.

Belanger, A.A. and Narayanan, R., 2008, Direct Strain Calculation of Pipe Line Dent from Knot Migration using a Kinematic Model Free of Material Properties. Proceedings of the International Pipeline Conference.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning.

Bitter, N.P. and Shepherd, J.E., 2013. On the Adequacy of Shell Models for Predicting Stresses and Strains in Thick-Walled Tubes Subjected to Detonation Loading. Pressure Vessels and Piping Conference, Paris, France.

Chen, G., and G. J. Hay. 2011. "A Support Vector Regression Approach to Estimate Forest Biophysical Parameters at the Object Level Using Airborne Lidar Transects and QuickBird Data."Photogrammetric Engineering & Remote Sensing 77: 733–741. doi:10.14358/PERS.77.7.733.

Ciarlet, P.G. and Mardare, C., 2008. An Introduction to Shell Theory. Differential Geometry: Theory and Applications, 9, pp.94-184.

Cosham, A. and Hopkins, P., 2003. The Pipeline Defect Assessment Manual (PDAM) – A Report to the PDAM Joint Industry Project. Newcastle, UK.

Cosham, A. and Hopkins, P., 2004, The Effect of Dents in Pipelines - Guidance in the Pipeline Defect Assessment Manual. International Journal of Pressure Vessels and Piping, 81(2), pp.127- 139.

CSA Z662, Oil and Gas Pipeline Systems 2016.

Dawson, S.J., Russell, A. and Patterson, A., 2006. Emerging Techniques for Enhanced Assessment and Analysis of Dents. Proceedings of the International Pipeline Conference, Calgary, Alberta, Canada.

Dinovitzer, A., Lazor, R., Carroll, L.B., Zhou, J., McCarver, F., Ironside, S., Raghu, D. and Keith, K., 2002, Geometric Dent Characterization. 4th International Pipeline Conference (pp. 1589-1598). Calgary, Canada.

Erickson, A., 2010. Fatigue Crack Failure Associated with Shallow Dents on Pipelines. NEB. File of-Surv-Inc-02.

Foroughi, H., Moen, C.D., Myers, A., Tootkaboni, M., Vieira, L. and Schafer, B.W., 2014. Analysis and Design of Thin Metallic Shell Structural Members-Current Practice and Future

Gao, M., & Krishnamurthy, R., 2015. Mechanical Damage in Pipelines: A Review of the Methods and Improvements in Characterization, Evaluation, and Mitigation. Oil and Gas Pipelines: Integrity and Safety Handbook, First Edition. Chapter 22.

Research Needs. in Proc. of Annual Stability Conference Structural Stability Research Council, Toronto, Canada.

Fowler, J.R., 1993, Criteria for Dent Acceptability in Offshore Pipeline. Offshore Technology Conference. Offshore Technology Conference, Houston, Texas, USA.

Gao, M., McNealy, R., Krishnamurthy, R. and Colquhoun, I., 2008. Strain-Based Models for Dent Assessment—A Review. ASME Paper No. IPC2008-64565. Proceedings of the International Pipeline Conference, Paper No. ASME, IPC04-0061, Calgary, Alberta, Canada.

Ghaednia, H., Das, S., Wang, R. and Kania, R., 2015. Safe Burst Strength of a Pipeline with Dent–Crack Defect: Effect of Crack Depth and Operating Pressure. Engineering Failure Analysis, 55, pp.288-299.

Hanif, W. and Kenny, S., 2014, Mechanical Damage and Fatigue Assessment of Dented Pipelines using FEA. Proceedings of the 10th International Pipeline Conference, Calgary, Canada.

Hojjati, M.H. and Lukasiewicz, S.A., 2008. Filtering Algorithm for Radial Displacement Measurements of a Dented Pipe. International Journal of Pressure Vessels and Piping, 85(5), pp.344-349.

Ironside, S.D. and Carroll, L.B., 2002, Pipeline Dent Management Program 2002 4th International Pipeline Conference (pp. 1859-1864). Calgary, Canada.

Karamanos, S.A. and Andreadakis, K.P., 2006. Denting of Internally Pressurized Tubes under Lateral Loads. International Journal of Mechanical Sciences, 48(10), pp.1080-1094.

Kocijan, J., Likar, B., 2008. Gas–liquid separator modelling and simulation with Gaussian-process models, Simulation Modelling Practice and Theory, 16(8), pp. 910-922

Koiter, W.T., 1970. The Stability of Elastic Equilibrium. Stanford Univ Ca Dept. of Aeronautics and Astronautics.

Leis, B.N., Forte, T.P. and Zhu, X., 2004. Integrity Analysis for Dents in Pipelines. Proceedings of the International Pipeline Conference, Paper No. ASME, IPC04-0061, Calgary, Alberta, Canada.

Love, A.E.H., 1888. The Small Free Vibrations and Deformation of a Thin Elastic Shell. Philosophical Transactions of the Royal Society of London. A, 179, pp.491-546.

Lukasiewicz, S.A., Czyz, J.A., Sun, C. and Adeeb, S., 2006, Calculation of Strains in Dents Based on High Resolution In-line Caliper Survey. Proceedings of the International Pipeline Conference, Calgary, Alberta, Canada, (pp. 129-134).

Luo, P.F. and Chen, J.N., 2000. Measurement of Curved-Surface Deformation in Cylindrical Coordinates. Experimental Mechanics, 40(4), pp.345-350.

Macdonald, K.A. and Cosham, A., 2005. Best Practice for the Assessment of Defects in Pipelines – Gouges and Dents. Engineering Failure Analysis, 12(5), pp.720-745.

Naghdi, P.M. and Nordgren, R.P., 1963. On the Nonlinear Theory of Elastic Shells under the Kirchhoff Hypothesis. Quarterly of Applied Mathematics, 21(1), pp.49-59.

Noronha, D.B., Martins, R.R., Jacob, B.P. and de Souza, E., 2010, Procedures for the Strain Based Assessment of Pipeline Dents. International Journal of Pressure Vessels and Piping, 87(5), pp.254-265.

Noronha, D.B., Martins, R.R., Jacob, B.P. and Souza, E., 2005. The use of B-Splines in the Assessment of Strain Levels Associated with Plain Dents. Proceedings of the Rio Pipeline Conference and Exposition.

Okoloekwe, C., Muntaseer, K., Langer, D., Hassanien, S., Cheng, R., and Adeeb, S. 2017. Deformation Analysis of Dented Pipelines via Surface Interpolation. Proceedings of the Pressure Vessels and Piping Conference, Hawaii, USA.

Ong, L.S., Soh, A.K. and Ong, J.H., 1992. Experimental and Finite Element Investigation of a Local Dent on a Pressurized Pipe. The Journal of Strain Analysis for Engineering Design, 27(3), pp.177-185.

Oshana Jajo, Jandark, 2014. Dent behaviour of steel pipes under pressure load. Electronic Theses and Dissertations. 5025.

Panetta, P.D., Diaz, A.A., Pappas, R.A., Taylor, T.T., Francini, R.B. and Johnson, K.I., 2001. Mechanical Damage Characterization in Pipelines. Pacific Northwest National Lab. Richland WASA35467

Race, J.M., Haswell, J.V., Owen, R. and Dalus, B., 2010. UKOPA Dent Assessment Algorithms: A Strategy for Prioritising Pipeline Dents. Proceedings of the 8th International Pipeline Conference, Calgary, Alberta, Canada.

Rafi, A.N.M., Das, S., Ghaednia, H., Silva, J., Kania, R. and Wang, R., 2012. Revisiting ASME Strain-Based Dent Evaluation Criterion. Journal of Pressure Vessel Technology, 134(4), p.041101.

Rafi, A.N.M., Das, S., Ghaednia, H., Silva, J., Kania, R. and Wang, R., 2012. Revisiting ASME Strain-Based Dent Evaluation Criterion. Journal of Pressure Vessel Technology, 134(4).

Reissner, E., 1952. Stress-Strain Relations in the Theory of Thin Elastic Shells. Studies in Applied Mathematics, 31(1-4), pp.109-119.

Rasmussen, C. E and Williams, C. K. I., 2005. Gaussian Processes for Machine Learning. The MIT Press.

Rasmussen, C. E and Nickisch, H., 2016. The GPML Toolbox version 4.2

Rogers, D.F. and Adams, J.A., 1990. Mathematical Elements for Computer Graphics, McGraw- Hill Book Co. New York.

Rosenfeld, M.J., Porter, P.C. and Cox, J.A., 1998, "Strain Estimation using Vetco Deformation Tool Data," Proceedings of the International Pipeline Conference, Calgary, Alberta, Canada.

Rosenfeld, M.J., Porter, P.C. and Cox, J.A., 1998. Strain Estimation using Vetco Deformation Tool Data. International Pipeline Conference, Calgary, Alberta, Canada.

Sanders Jr, J.L., 1963. Nonlinear Theories for Thin Shells. Quarterly of Applied Mathematics, 21(1), pp.21-36.

Ventsel, E. and Krauthammer, T., 2001. Thin Plates and Shells: Theory: Analysis and Applications. CRC press.

Woo, J., Muntaseer, K. and Adeeb, S. 2017. Development of A Profile Matching Criteria to Model Dents in Pipelines using Finite Element Analysis Pressure Vessels and Piping Conference, Hawaii, USA.

Wu, Y.Z.P. and Han, X., 2013. Analysis of Pipe Size Influence on Pipeline Displacement with Plain Dent Based on FE Calculation. International Journal of Computer Science Issues, 10(1), pp.507-510.

Zhao, K., S. Popescu, X. Meng, Y. Pang, and M. Agca. 2011. "Characterizing Forest Canopy Structure with Lidar Composite Metrics and Machine Learning." Remote Sensing of Environment 115: 1978–1996. doi:10.1016/j.rse.2011.04.001.

Zhao, K., S. Popescu, and X. Zhang. 2008. "Bayesian Learning with Gaussian Processes for Supervised Classification of Hyperspectral Data." Photogrammetric Engineering & Remote Sensing 74: 1223–1234. doi:10.14358/PERS.74.10.1223.