arXiv:2611.30032v1 [stat.AP] 23 Nov 2026

基于模型的基于图像的人工检测检出概率研究

Qizheng Xia * William Q. Meeker † Joseph N. Gray ‡ Qing Li §

2026年11月22日

摘要

确定老化系统中组件可靠性的一项重要工作，是为所采用的检测方法建立检出概率（POD）曲线。传统上，检出概率曲线是根据一组含有代表性缺陷的实物试样，以及记录检出结果的检测人员进行估计的。这类实验可能成本高昂，尤其是因为制造具有真实缺陷特征的试样需要较高成本。

本文展示了经过适当验证的基于物理的仿真如何大幅降低成本，并在适用的情况下有望替代昂贵的实体检出概率实验。无损评价（NDE）仿真提供了一种替代方案，能够以较低成本生成图像，同时呈现比实体实验通常可行的范围更广的缺陷形态、缺陷位置及其他检测变异性来源。在本研究中，我们使用XRSIM（一种X射线射线照相模拟器），研究对含有不同尺寸和位置缺陷的焊缝进行检测时的检出概率。缺陷特征根据指定分布随机生成，缺陷位置和每幅图像中的缺陷数量也经过随机化；每幅图像包含零至三个缺陷。每个独立的模拟缺陷均作为一个观测单位，因此包含多个缺陷的图像会贡献多个缺陷层面的命中－漏检响应。

为纳入人工检测变异性，将模拟图像以随机顺序呈现给五名检验员，每名检验员对这些图像检测两次。所得的命中-未命中数据采用重复测量二元回归模型进行分析，该模型考虑了缺陷之间、检验员之间、缺陷/检验员组合之间以及检测之间的变异性。随后，利用拟合模型估计总体均值POD曲线及其相应的统计不确定性。结果表明，缺陷间变异占主导地位，同时存在明显的检测人员间差异。

关键词：贝叶斯估计、模型辅助检出概率、MAPOD、检出概率、射线照相、重复测量、X射线仿真、XRSIM。

*工业与制造系统工程系，爱荷华州立大学，艾姆斯，IA 50010，美国

†统计系，爱荷华州立大学，艾姆斯，IA 50010，美国

‡无损检测中心，爱荷华州立大学，艾姆斯，IA50010，美国，§工业与制造系统工程系，无损检测中心，爱荷华州立大学，艾姆斯，IA50010，美国，qlijane@iastate.edu（通讯作者）

1 引言

1.1 背景与动机

实验检出概率测量旨在对典型检测中的无损检测信号变异性进行抽样。无损检测中检验的变异来源包括仪器变异性、设置变异性、缺陷形态变异性、人类感知变异性、数据查看变异性（尤其是灰度图像查看）等。制备这些检出概率实验测量值的成本，主要受以下因素制约：用于涵盖工件和缺陷形态检测几何特征的试样块数量有限、试验中可用的检验员人数有限，以及研究可用设备的范围与现场通常使用的设备可能范围相比有限。由于存在这些限制，而且人们怀疑经济可行的实验POD研究能否充分涵盖多参数空间，因此无损检测模拟器等替代方案颇具吸引力。

在本文中，我们使用“缺陷”这一通用术语来指代图像中需要检测的特征。在我们的应用中，这些缺陷是裂纹；在讨论这些缺陷时，我们使用更具体的术语“裂纹”。

1.2 基于图像的人工检测面临的挑战

本研究中使用无损检测模拟器的示例，重点关注典型的X射线射线照相二维（2D）图像，并使用X射线仿真程序XRSIM（Xu 等人，1994；Jensen 和 Gray，1995；Elshafiey 和 Gray，1996；Jensen 等人，1996；Roth 等人，2011）。大多数无损检测测量都使用大型探头；这里所说的“大型”，是指通常检测到的缺陷只占探测能量作用区域的一小部分。这方面的例子包括超声检测，其中波束尺寸通常大于反射体；以及涡流检测，其中线圈产生的感应电流流动相对于产生指示的特征而言范围较大。在这类测量中，信号尺寸大致与缺陷尺寸相关。在X射线成像中，探测器像素尺寸通常远小于所关注缺陷的尺寸。事实上，如果缺陷尺寸小于一个像素，则该缺陷不会被视为可检测到。即使对于对比度非常高、对比噪声比（CNR）显著大于3:1的特征，情况也是如此。对于二维成像无损检测方法而言，像素（点测量）较小，即小于150微米，这一观察结果导致了在定义何种特征可被检测到方面的重要差异。在X 射线图像中，CNR小于1的缺陷也很容易被检测到。在图1 中，我们看到两个CNR分别为0.6和0.3的缺陷，此时图像有助于提高检出限。探测器分辨率，尤其是像素尺寸，在任何X射线检测中都十分重要。

在图 1中，我们还可以看到每幅图像的左侧都有一条穿过缺陷的线迹，显示背景噪声和缺陷对比度。在每幅图像的右侧，都可以清楚地看到缺陷。线迹清晰地表明，这些图像的 CNR 较低。利用数百个像素对缺陷进行这种多点测量，体现了人类感知系统识别噪声中模式的能力。噪声导致缺陷图像中的所有像素相对于背景均值产生相同的微小偏移的可能性非常小。其结果是，X射线图像的可探测性损失发生在对比度处于噪声水平时，而不是由点探测器所需的大幅 3σ 偏差所决定的。由噪声波动导致的误报非常少。

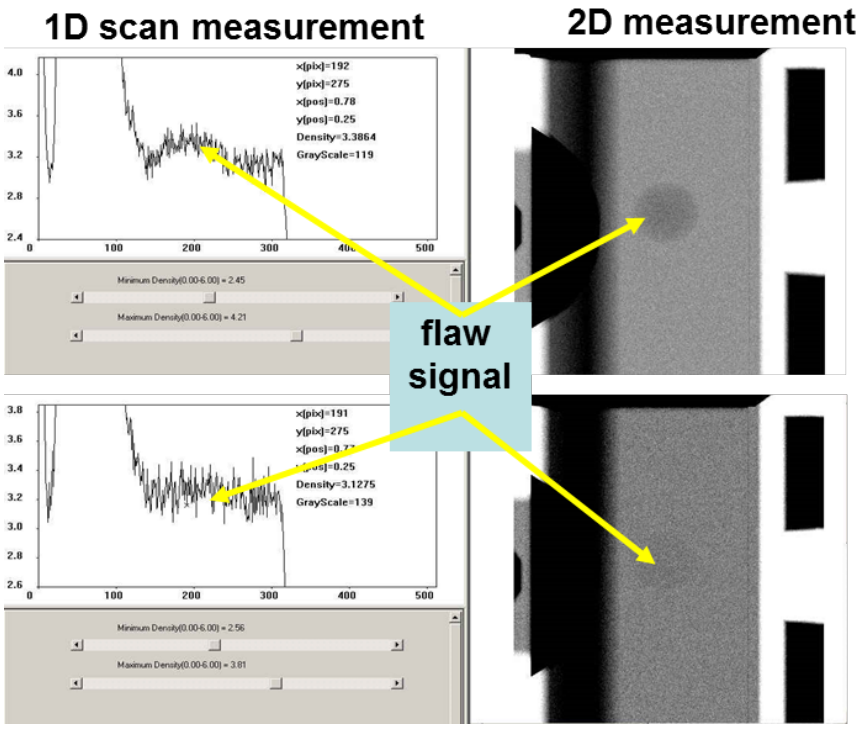


图1：右上方的图像的对比噪声比为 0.6:1，而右下方的图像的对比噪声比为 0.3:1。左侧的图表显示了穿过缺陷中心的水平线轨迹，清楚地表明对比噪声比较低。

在生成用于估计检出概率的数据集时，我们需要捕捉缺陷信号实验测量值的变异性，以及人类解读数据时的变异性。影响缺陷信号的因素包括缺陷尺寸和形态、缺陷在工件中的位置、工件的几何形状，以及测量系统的特性，例如探测器类型、X射线源和空间分辨率。此外，还需要考虑其他参数，包括数据查看条件、由工件制造或老化过程控制的初始缺陷分布，以及检验员的变异性，包括检验员所接受的培训。许多基于二维数据集的检测无法像通常所采用的方式那样简化为一个参数，例如将检出概率建模为裂纹长度的函数。数据的二维特性提供了比单点测量更多的维度信息。最后，一些检测中，性能退化与缺陷的某个单一参数之间存在直接关系，而另一些检测则具有更复杂的关系。对于裂纹而言，裂纹长度和穿透材料的深度都会影响结构完整性，尽管X射线检测的概率可能由其他缺陷特征决定。

利用无损检测仿真程序生成典型的合成无损检测数据时，我们有机会生成涵盖各种检测变量的数据，而由于成本限制，这些变量过去通常未被纳入POD研究。最典型的例子是部件形状的复杂性，尤其是缺陷形态的复杂性。在任意位置制造具有实际部件典型形态的缺陷，成本高得令人难以承受。即使是在若干样品中设置数百个缺陷这样相对适度的范围，成本也很高。这些样品的有限可用性体现在：一旦材料的合金发生变化，通常就需要针对新材料开展一项新的研究。如果部件上的载荷发生变化，从而产生不同的缺陷总体，则还需要开展新的研究并制作新的样品。

1.3 相关工作

POD提供了一种统计度量，用于衡量检测程序检测缺陷的能力。传统的POD研究会估计检测程序检测到缺陷的概率，并将其表示为缺陷尺寸或其他相关缺陷参数的函数。常见的命中/未命中方法和信号-响应方法之所以有用，是因为它们将检测性能归结为诸如 a_{90} （估计的检出概率为0.90时的缺陷尺寸，或等价地，漏检概率为0.10）和$a_{90/95}$ （ a_{90}的上限95%置信界）等量，但它们还需要足够多的缺陷样品来体现检测的变异性。模型辅助探测概率（MAPOD）旨在通过将经验观测结果与基于物理的模型信息相结合，减少这种实验负担。Li等人（2014年）在无损检测中使用物理模型进行POD估计，并展示了如何利用模型信息补充有限的实验数据。Le Gratiet等人（2017）提出了渐进式模型辅助探测概率方法，将仿真、不确定性量化和POD估计联系在同一个分析过程中。Bismut和Straub（2022）对概率型无损检测模型进行了综述，并强调应结合检测决策以及影响测量的不确定性来源来解读POD。这些研究为在POD研究中使用仿真提供了一般方法学基础，但没有涉及X射线射线照相中出现的具体成像和人工观察问题。

多项研究通过对X射线过程本身建模，探讨了X射线检测能力。Koshti(2014年)提出了一种用于检出概率研究的X射线缺陷尺寸参数，将裂纹几何形状和检测设置因素整合为一个参数，旨在使其与X射线图像对比度及检出概率呈单调相关。该研究指出，射线照相中的裂纹可检测性并非仅由某一个

物理尺寸决定；X射线角度、探测器几何结构、裂纹开口、部件厚度和探测器响应都可能改变有效指示。在后续研究中，Koshti (2015)利用模拟对比度和信噪比计算，辅助制定能够实现预期缺陷可检出性的射线照相技术，并展示了裂纹深度、裂纹宽度、探测器分辨率、源尺寸和几何结构的变化如何影响预测的X射线缺陷参数和图像对比度。这两种模型均采用板材中的矩形槽，并依赖响应校准，同时假定散射较低、均匀且受控。Koshti (2017)还将这种建模扩展至掠射角X射线，并指出，沿裂纹表面方向入射的低能X射线可通过光滑、平坦裂纹表面之间的镜面反射改变裂纹对比度。这种效应的适用性受低X射线能量下材料衰减的限制。不过，这些研究主要是定义或模拟X射线缺陷参数，并未针对模拟射线照片收集多次人工命中/未命中数据。

1.4 拟议研究概述

我们采用模型辅助概率检出框架（MAPOD）开展基于图像的射线照相人工检测研究。本研究使用 XRSIM 生成合成焊缝缺陷图像，并对缺陷尺寸、缺陷形态、缺陷位置、相对于入射 X 射线束的缺陷方向以及典型焊缝射束厚度进行受控变化。不同于仅依赖仿真或开展自动图像分析研究的方法，本研究在受控观察协议下由人工观察者完成最终检测响应，从而能够在概率检出分析中体现检验员变异性和观察效应。所得命中-漏检数据分别采用传统的单次检测二元回归模型和重复检测贝叶斯随机效应模型进行分析，后者考虑了缺陷间、检验员间以及重复检测的变异性。通过将基于物理的图像仿真与受控人工检测相结合，本研究表明，模拟射线照片能够在不完全依赖昂贵实体缺陷试样的情况下，显著降低POD研究的实验负担。

1.5 概述

本文其余部分安排如下。第2节介绍模拟图像的生成、缺陷参数的选择以及检验/观察协议。第3节介绍命中-漏检POD数据的统计模型，并说明如何使用这些模型估计POD。第4节介绍检测实验、所得的命中/未命中数据以及POD估计结果。第5节讨论对MAPOD研究的启示，以及未来相关研究的方向。

2 模型辅助POD方法：模拟X射线图像生成与人工检测分析

本节介绍生成检验数据所采用的方法。XRSIM 用于生成焊缝的合成射线照相图像，这些图像具有逼真的焊道几何形状，且其中的缺陷特征、缺陷位置和材料厚度条件均经过可控变化。该过程包括选择缺陷类型、定义相关缺陷参数、随机化缺陷几何形状和位置，以及指定图像阅片方案，其中包括灰度映射和检验覆盖范围。第3节概述了适用于重复测量检出概率

数据的统计方法，第4.3 节则将这些方法用于估计检出概率。这些步骤共同构成了一个模型辅助框架，将基于物理的图像仿真与人工检测相结合，以进行检出概率估计。本研究的整体流程如图2所示。该流程首先生成缺陷参数和位置均受控的模拟X射线图像，随后收集重复的人工检出／漏检检测响应，最后使用统计检出概率模型估计检出概率曲线。

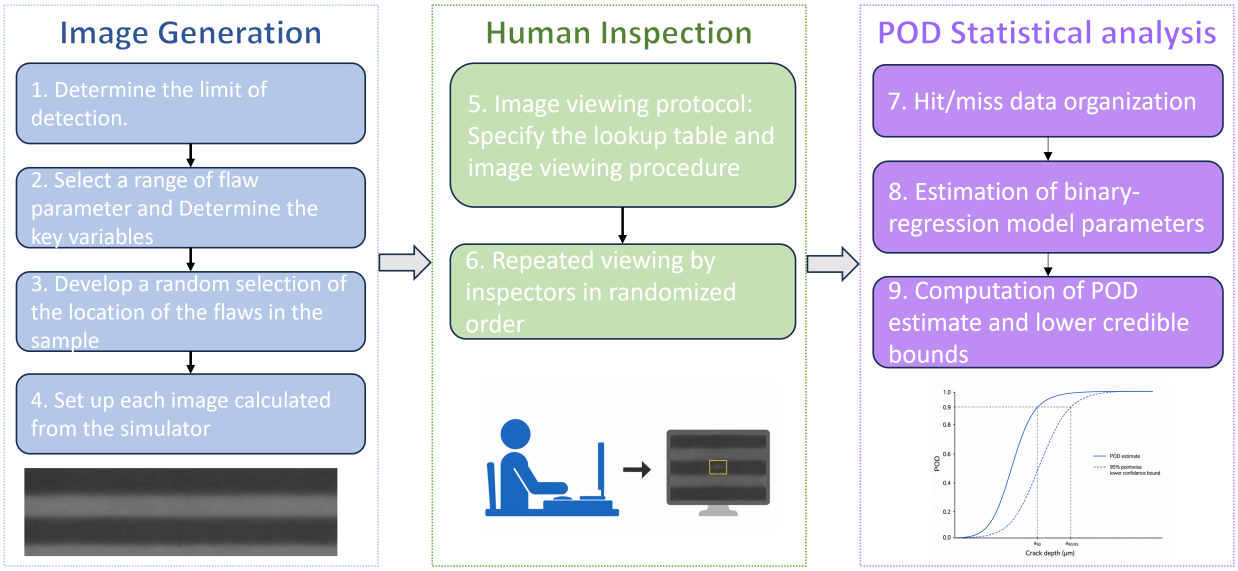


图 2：所提出的基于图像的 X 射线人工检测模型辅助检出概率研究的工作流程。

2.1 基于仿真的图像生成

POD研究中的X射线图像生成，首先需要确定工件中材料厚度范围对应的可检测性极限。具体做法是使用一个包含32个阶梯的阶梯楔块，其最大厚度和最小阶梯厚度分别界定工件中的厚度范围。对于每个阶梯厚度，布置一组缺陷尺寸，范围从容易检测的大缺陷到低于对比度检测限的极小缺陷。该检测限由成像过程中的噪声决定，见图 1。由于需要以该缺陷尺寸为中心进行更多测试，并对容易观察到的缺陷和在另一端完全无法观察到的缺陷进行采样，因此这一可检测性极限十分重要。利用计算机辅助设计（CAD）模型，在XRSIM中生成的厚度图可以轻松确定厚度范围，更重要的是，还能确定该变化发生的位置，如图3所示。

2.2 缺陷类型和缺陷参数设计

接下来需要解决的问题是初始缺陷分布中缺陷的范围。在焊接检测中，缺陷分为几类，包括裂纹、气孔和夹杂物。后两者属于体积型缺陷。裂纹由于其宽度尺寸相对于另外两个垂直方向的尺寸非常小，因此具有很强的方向依赖性。裂纹缺陷之所以

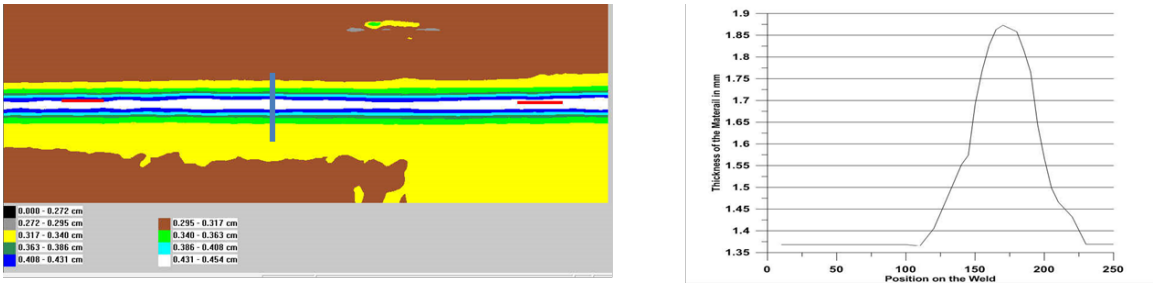


图 3：左图是用于缺陷可探测性研究的焊缝厚度图。颜色标尺表示 X射线投影厚度。右侧曲线图显示的是厚度图中蓝线所在位置的厚度。该焊缝连接两块板材。两条水平红线表示焊缝材料中厚度不同位置的两个缺陷。

被选用于本研究，是因为采用X射线射线照相技术检测此类缺陷十分困难。影响裂纹能否被检出的参数很多，包括决定缺陷信号的裂纹深度和宽度。裂纹相对于入射X射线束的取向是一个非常敏感的参数，因为对于非常紧闭的裂纹，其取向只需改变几度，缺陷信号就会从易于检出的信号变成没有对比度、因而无法检出的信号。这一点很重要，因为它表明，了解缺陷分布与检出概率中的人为因素变化同样重要。因检测信号消失而漏检缺陷，与检验员在检测过程中暂时走神并不相同。

裂纹开口对 X射线检测至关重要。因此，本文仅考虑贯穿表面的、无应力的（张开）裂纹，并设置了多种裂纹张开宽度。对于生成的 105 个人工缺陷，图像中的裂纹位置、沿焊缝的位置，以及位于不同焊缝余高厚度区域内的位置均为随机选取。因此，模拟缺陷取样自一个假定的裂纹总体，该总体中的裂纹具有不同的方向、深度、宽度，并且在沿焊缝方向及不同焊缝余高厚度区域内具有不同位置。

由于工件厚度不均匀，因此需要进行两次曝光，因为对于薄区，适当的曝光会导致厚区曝光不足。而对于第二次曝光，针对厚区的最佳曝光会使工件的薄区饱和。对于数字探测器，任何曝光的最佳条件通常是最大位深度的 80%。因此，对于 16 位探测器，平均灰度应为 52,000 个灰度级。采用这一标准的原因在于，X射线图像中的主要噪声受采集到的光子总数控制。该过程遵循泊松噪声分布，其中噪声与探测器采集到的光子总数的平方根成正比。光子计数越大，泊松噪声越低。这种噪声决定了探测器的对比度灵敏度，具体而言，即可检测到的最小缺陷尺寸。

2.3 缺陷特征和位置的随机化

为生成图像，针对缺陷计算了随机数，并改变了裂纹深度、宽度、沿焊道的位置以及在焊缝厚度变化区域中的位置。将深度作为关键参数，以便把所有影响缺陷信号的变量简化为一个变量，用于命中/未命中图和检出概率图。裂纹深度相对于

厚度的灵敏度图用于在信噪比较低的深度范围内选取更多缺陷。对于每个缺陷，都对相应的裂纹 CAD 模型进行缩放，使其符合从初始分布中随机选取的缺陷参数。随后，将缺陷的 CAD 表示置于焊道的 CAD 模型中焊缝内的适当位置。这需要指定缺陷包围盒相对于焊道的（x，y，{v7>）坐标。

作为比较，在焊缝中的随机位置制造 100+ 真实物理缺陷，成本将极其高昂。尽管使用 CAD模型和 XRSIM接口生成数量相同的合成裂纹缺陷耗时较长，但采用模拟的人工缺陷是可行的。需要指出的是，使用三维计算机断层扫描对具有粗糙表面的真实裂纹进行表征，可以纳入一定程度的缺陷保真度，而在试样中制造的缺陷无法达到这种程度。

2.4 图像查看协议

图像示例见图4。这些示例选自含有较大缺陷的图像；由于计算机显示器显示的 8位图像难以充分呈现 14位或 16位图像，因此这些较大缺陷在打印件上更容易观察到。这一限制带来了若干问题，因此在制定检测协议以及培训检验员查看图像时，都需要格外谨慎（例如，Kimpe andTuytschaever，2007；Barten，1992，1999）。

首先，考虑16位图像的查看问题。现代数字X射线探测器的噪声水平通常为50–70个灰度级，而16位饱和级别约为65,000个灰度级。因此，将16位灰度映射到8位灰度（这是显示器显示能力的上限）是影响检测人员表现的一个关键因素。具体而言，必须针对工件中的不同厚度范围指定适当的映射，因为灰度会随着焊缝厚度的变化而改变。当焊缝厚度跨越超过6000个灰度级，而缺陷检测灵敏度约为70个灰度级时，用于检查图像的查找表（LUT）必须作为检测协议的一部分加以规定。

在图 5中，A面板针对焊缝较厚部分设置了16位到8位的灰度映射（由特定的查找表定义），并且具有非常窄的灰度范围，适合查看。B面板对灰度范围的容许度更高，但同样显示了查找表的影响。在两个面板中，适用于一种厚度的查找表都会使查看另一种厚度成为不可能，因此需要使用多个映射来查看图像。请注意，每个面板中的光标都指向裂纹缺陷。

观察图像时的第二个问题是，确保系统地检查整个图像。研究表明，观察者往往会在图像中的显著特征之间移动视线，而不是均匀地检查所有区域（例如，Livingstone 和 Hubel，2008）。眼动追踪研究通过摄像机记录观察者观看图像时的视线，证实了这一倾向。这一问题在医学成像中尤为值得关注，因为放射科医师经常需要检查大型图像。一种促使观察者全面覆盖图像的方法是使用移动掩膜，系统地逐块扫描图像中的较小区域，要求观察者逐区域检查图像。为此，我们开发了一种简单的掩膜。

2.5 生成图像的步骤总结

总而言之，为检查员命中-漏检评估生成图像所需的步骤如下：



图4：图中显示了位于相同位置、深度范围为80微米至200微米的四幅宽裂纹图像。该裂纹的所有参数均大于测试图像中所使用的参数。尺寸的缩小体现了缺陷相对于背景的对比度降低。需要注意的是，这些缺陷位于焊缝的平板部分，在此位置更容易发现裂纹。

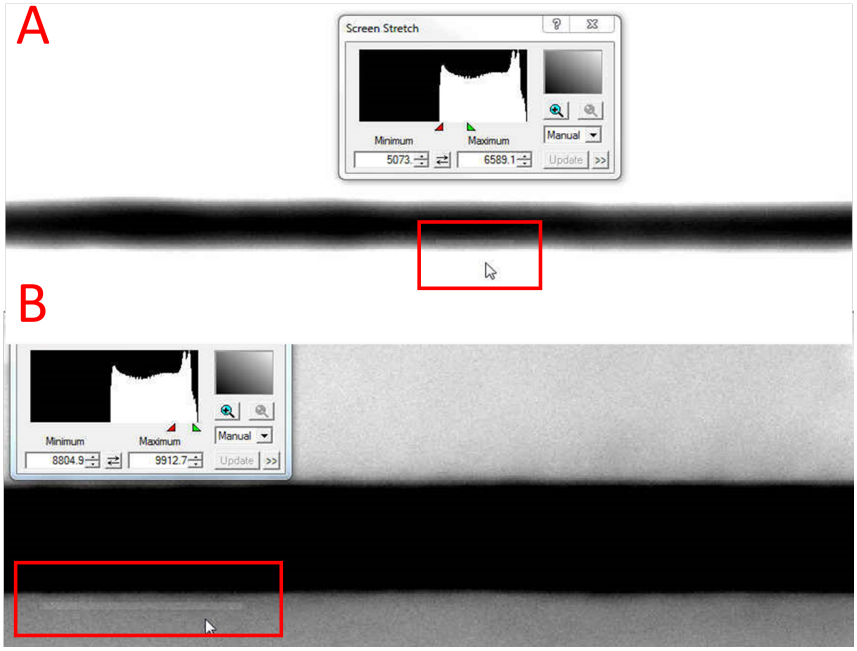


图5：面板 A 和 B 显示了焊道中不同位置的两个焊缝缺陷，需要采用不同的映射将 16 位数据转换为 8 位显示器。小的红色和绿色三角形显示了映射到 8 位显示器灰度的 16 位数据范围。每个面板中的光标指向一个裂纹缺陷。

- 使用仿真工具确定检出限，以及缺陷相对于 X 射线投影方向的方向。这必须覆盖样品中厚度变化的范围。

- 选择一组能够代表检测中预期初始缺陷分布的缺陷参数范围，并围绕检出限对这些缺陷进行采样。在本研究中，我们选择了裂纹的缺口表征，这类裂纹通常可用于实验测定检出概率。

- 确定影响缺陷信号的关键变量，在本例中包括裂纹开口、管电压以及裂纹在焊道中的位置。

- 随机选择缺陷在样品中的位置。在本例中，我们改变了缺陷横跨焊道和沿焊道方向的位置。一幅图像中的缺陷数量从零个到三个不等。
- 使用模拟器生成每幅图像，并记录缺陷的实际位置。
- 为图像制定适当的阅片方案，明确用于将 16 位探测器数据映射到 8 位计算机显示器的查找表，并制定光栅扫描掩模，以确保查看整幅图像。

- 建立一种方法，跟踪与检测图像顺序随机化相关的记录工作。我们使用包含所有关键参数的文件名，并建立映射，将其对应到第二组文件名；这组文件名包含用于指定阅片顺序的数字。每次检测图像时，我们都会重新生成阅片顺序。

- 将检测结果记录为命中-漏检数据，注明每张图像中存在的哪些缺陷被发现，哪些缺陷被漏检。

3 命中-漏检POD数据的统计模型

3.1 单次检测命中-漏检POD数据的统计模型

在一项POD研究中，如果可用的缺陷（在本例中为图像中的人工缺陷）仅接受一次检测，通常会使用简单的二元回归模型来分析数据。该模型的详细描述见MIL-HDBK-1823A（2009，附录 G），此处不再重复。我们仅回顾基本概念。二元随机变量，

$$H = \begin{cases} 1, & \text{if the flaw is detected (a hit)}, \\ 0, & \text{if the flaw is not detected (a miss)} \end{cases}$$

是响应变量。令 x 表示缺陷的一个适当特征（可能经过变换），例如图像中的裂纹长度或缺陷面积，该特征对检出概率有重要影响。检出概率的二元回归模型为

$$\mathrm{POD}(x) = \Pr(H = 1|x) = \Phi(\beta_0 + \beta_1 x), \tag{1}$$

其中，$\Phi(\cdot)$ 是位置-尺度（LS）分布的标准累积分布函数。常用的LS分布包括正态分布、逻辑斯蒂分布、最小极值分布和最大极值分布。在二元回归的广义线性模型（GLM）表述中，所选分布分别对应于连接函数（概率单位、逻辑斯蒂、互补对数-对数和对数-对数）的选择。这些LS分布的标准累积分布函数分别记为 Φ_{norm}、Φ_{logis}、Φ_{sev}和Φ_{lev}。Meeker等人（2022，第4章和附录 B.3）详细介绍了这些分布及其他LS分布。

对模型（1）进行重新参数化很有用，方法是使用

$$\mathrm{POD}(x) = \Pr(H = 1|x) = \Phi\left(\frac{x - \mu}{\sigma}\right). \tag{2}$$

对于 $\beta_1 > 0$，这意味着检出概率曲线递增，$\mu = -\beta_0/\beta_1$ 和 $\sigma = 1/\beta_1$。这种重新参数化的重要优势在于，μ 和 σ 与缺陷尺寸 x 具有相同的单位，且具有实用的解释。具体而言，μ 是描述在特定检测中恰好能够被检出的（随机）缺陷尺寸的 LS分布的位置参数，而 σ 是相应的尺度参数，用于描述检出概率曲线随 x增加的速度。对于对称分布（例如正态分布和逻辑斯蒂分布），μ是均值/中位数；对于正态分布，σ是标准差。采用（μ，σ）参数化的另一个优势是，这些参数是稳定参数；与（β_0，β_1）不同，它们通常更容易识别，相关性也较低。

目前，二元数据最常用的统计估计方法是最大似然法（ML）。最大似然法通过寻找使数据出现概率最大的参数值来进行估计。关于将最大似然法用于检出概率估计的详细信息，请参阅MIL-HDBK-1823A (2009，附录 G），此处不再赘述。统计不确定性（即由数据有限造成的不确定性）通常通过报告估计值的标准误来量化

或置信区间。 MIL-HDBK-1823A(2009，附录 G）还介绍了进行相关计算的方法。

传统上，报告检出概率函数的点估计时，也会一并报告一组逐点下置信界（有时称为置信带）。计算置信界限的方法有多种，MIL-HDBK-1823A(2009，附录 G）中也对此作了介绍。尽管基于似然的置信区间方法在理论上具有更好的性质，但易于计算的“Wald”方法最为常用；对于无损检测应用中通常采用的样本量（例如，使用命中/未命中数据进行研究时，缺陷数量达到 60 个或更多），该方法通常能提供足够准确的近似，前提是缺陷尺寸范围内的重要部分没有数据稀疏的情况。

3.2 重复测量命中-未命中检出概率数据的统计模型 a

3.2.1 背景

第3.1 节概述的相对简单的统计模型有一项重要局限：它不够通用，无法处理对缺陷进行多次检测的实验。例如，如果有 100 个缺陷可供检测，并且四名不同的检测人员都检测了所有缺陷，那么所获得的信息量与一名检验员检测 400 个不同缺陷时并不相同。要从重复测量实验中估计检出概率，需要采用特殊统计方法。此前，a 与 a检出概率数据的此类方法由Li等人(2014)针对命中-未命中检出概率数据，以及由Koh和Meeker(2014、2019)进行了介绍。

为了正确分析重复测量 POD 数据，模型中必须纳入随机效应，以描述实验数据中的其他变异来源，例如检验员差异、缺陷形态差异以及实验设置差异。

3.2.2 实验数据结构与符号

考虑一项实验产生的命中-未命中 POD 数据，其中缺陷与检验员采用交叉设计。缺陷i 的已知尺寸为 x_i，$i = 1$、.....、I，检验员 j，$j = 1$、.....、J对这些缺陷进行检验，并且可以对同一缺陷/检验员组合进行重复检验。令 k，$k = 1$、.....、K_{ij} 表示嵌套在缺陷/检验员组合内的重复次数（i，j）。在平衡实验中，每位检验员都会检验所有缺陷，并且 K_{ij} 对于所有 i 和 j组合而言都相同。图 6 展示了实验设计及其产生的数据。

3.2.3 潜在随机变量模型公式

为便于解释二元回归模型，我们引入一个不可观测的（有时称为潜在的）连续随机变量，该变量可解释为在特定检测中刚好能够被检测到的缺陷尺寸：

$$Y_{ijk} = \mu + \tau_i + \delta_j + \gamma_{ij} + \epsilon_{ijk}, \tag{3}$$

其中

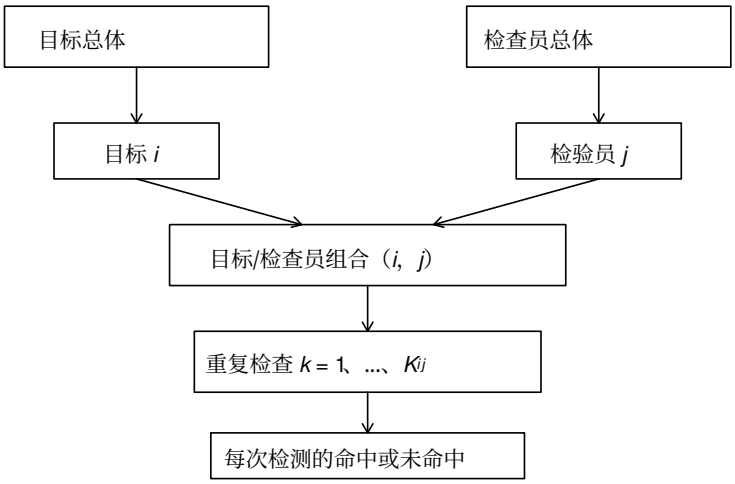


图6：实验数据结构示意图，其中缺陷与检测人员交叉，重复检查嵌套于缺陷/检验员组合内。

$\tau_i \sim$ 正态分布(0，σ_τ)　　是缺陷间随机效应，例如由随机裂纹形态引起。

$\delta_j \sim$ 正态分布(0，σ_δ)　　是检验员间随机效应，例如由经验和技能引起。

$\gamma_{ij} \sim$ 正态分布(0，σ_γ)　　是一个随机的缺陷/检验员组合效应，例如，检验员 j 可能难以检测缺陷 i。

μ　　是 Y_{ijk} 在$\tau_i = \delta_j = \gamma_{ij} = 0$条件下分布的位置参数。

$\epsilon_{ijk} \sim$ 位置-尺度$(0, \sigma)$　　是一个随机误差项，服从LS分布，用于描述检验员 j对缺陷 i进行重复检验时额外的检测间变异性。

图7 示意图　　统计模型的层级结构 Y_{ijk}。由（3）可知

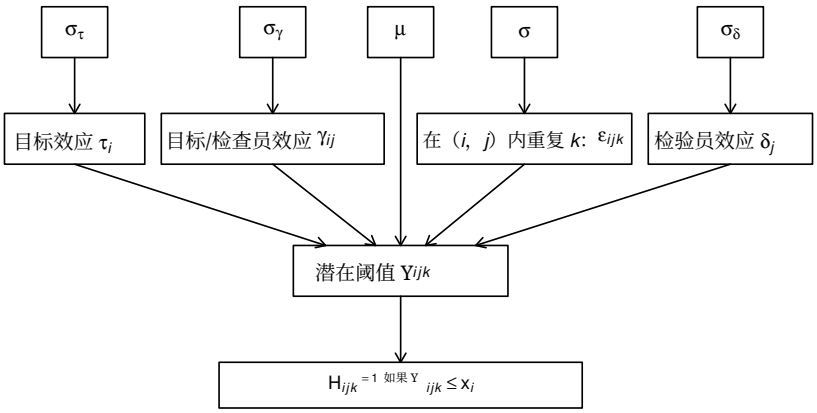


图7：潜在阈值 Y_{ijk} 及其变异来源的分层统计模型示意图。

任何随机效应的增加都会使 Y_{ijk} 增大，进而意味着缺陷更难检测。

同一检验员对同一缺陷进行重复检查时，τ_i、δ_j和γ_{ij}的取值相同。此外，这些随机效应与ϵ_{ijk} 在其所有索引取值之间彼此独立。随机效应标准差 σ_τ、σ_δ和σ_γ 分别描述缺陷、检验员以及缺陷/检验员组合的变异性。随机变量 ϵ_{ijk}/σ 服从标准 LS 分布；因此，

$$\Pr\left(\frac{\epsilon_{ijk}}{\sigma} \le z\right) = \Phi(z). \tag{4}$$

注意，Y_{ijk} 以及全部五个总体层面参数 μ、σ、σ_τ、σ_δ和σ_γ 均与缺陷尺寸具有相同的物理单位。与通常的广义线性模型二元回归公式相比，这一特性具有重要优势。

回想一下，一幅图像中的缺陷数量范围为零至三个。同一图像中多个缺陷的命中-漏检响应可能存在相关性，因为它们具有相同的图像层面条件。然而，由于对模拟图像背景和观察条件进行了严格控制，我们预计这种依赖性相对于显著的缺陷间变异而言很小，因此不对图像层面效应进行显式建模。

3.2.4 命中的条件概率

根据潜变量 Y_{ijk} 的定义，当 $Y_{ijk} \le x_i$ 时即为命中。利用（3）和（4）中的结果，在随机效应 τ_i、δ_j和 γ_{ij}的固定值条件下，命中的概率为

$$\begin{aligned} p_{ij} = \Pr(H_{ijk} = 1 \mid \tau_i, \delta_j, \gamma_{ij}) &= \Pr\left(Y_{ijk} \le x_i \mid \tau_i, \delta_j, \gamma_{ij}\right) \\ &= \Pr\left[\epsilon_{ijk} \le x_i - (\mu + \tau_i + \delta_j + \gamma_{ij}) \mid \tau_i, \delta_j, \gamma_{ij}\right] \\ &= \Phi\left[\frac{x_i - (\mu + \tau_i + \delta_j + \gamma_{ij})}{\sigma}\right]. \end{aligned} \tag{5}$$

注意，p_{ij} 不依赖于 k，因此检验员 j 对缺陷 i 进行的 K_{ij} 次重复检验具有相同的条件命中概率。

3.2.5 条件伯努利似然

现代贝叶斯软件（本文使用该软件进行估计）要求指定数据的似然，即在分层统计模型中给定任意随机效应值的条件下的数据似然。在随机效应的固定值条件下，

$$H_{ijk} \mid \tau_i, \delta_j, \gamma_{ij} \sim \text{Bernoulli}(p_{ij}), \tag{6}$$

其中p_{ij} 由（5）给出。因此，条件似然为

$$L_c(\mu, \sigma, \boldsymbol{\tau}, \boldsymbol{\delta}, \boldsymbol{\gamma}) = \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K_{ij}} p_{ij}^{h_{ijk}} (1 - p_{ij})^{1 - h_{ijk}},$$

其中 h_{ijk} 是 H_{ijk} 针对检验员 k 对缺陷 j 进行检测时的实际 0/1 值，i 而 $\boldsymbol{\tau}$、$\boldsymbol{\delta}$和$\boldsymbol{\gamma}$ 是包含其所有索引值对应随机效应的向量。于是，式（6）连同

针对 τ_i、 δ_j 和 γ_{ij} 的正态分布共同指定了层次模型。进行贝叶斯推断还需要总体层面参数的先验分布。随后将随机效应 τ_i、 δ_j 和γ_{ij} （对应 $i = 1$、.....、 I 以及 $j = 1$、.....、 J）与总体层面参数 μ、 σ、σ_τ、 σ_δ 和 σ_γ 联合估计。

3.2.6 条件POD

对于尺寸为 x的缺陷，其缺陷效应为 τ、检验员效应为 δ、缺陷/检验员组合效应为 γ时，单次检验的条件POD为

$$\mathrm{POD}(x \mid \tau, \delta, \gamma) = \Pr(Y \le x \mid \tau, \delta, \gamma) = \Phi\left[\frac{x - (\mu + \tau + \delta + \gamma)}{\sigma}\right].$$

请注意， $\mu+\tau+\delta+\gamma$是条件分布的位置参数，表示刚好能够检出的 Y（即缺陷尺寸）。对于正态分布和逻辑斯蒂分布等对称分布，该值也等于a_{50} （中位数）。与（2）中一样， σ 是尺度参数，决定检出概率曲线随 x增加的快慢。

3.2.7 总体均值（无条件）POD

对于总体均值POD，考虑分别从缺陷总体和检验员总体中随机选取一个缺陷和一名检验员，并由该检验员对该缺陷进行单次检验。此时，效应 τ、 δ和 γ 均为随机变量，必须对其分布求条件POD的均值。令$\psi = \tau + \delta + \gamma$。那么 $\psi \sim$服从正态分布~~Normal(0, σ_ψ)~~，其中 $\sigma_\psi = \sqrt{\sigma_\tau^2 + \sigma_\delta^2 + \sigma_\gamma^2}$，且

$$\mathrm{POD}(x) = \mathrm{E}\left[\Phi\left(\frac{x - (\mu + \psi)}{\sigma}\right)\right].$$

使用

$$\psi = \sigma_\psi Z, \qquad Z \sim \mathrm{Normal}(0, 1),$$

期望值使用.....计算

$$\mathrm{POD}(x) = \int_{-\infty}^{\infty} \Phi\left[\frac{x - (\mu + \sigma_\psi z)}{\sigma}\right] \phi_{\mathrm{norm}}(z)\, dz. \tag{7}$$

当随机误差项 $\epsilon_{ijk} \sim$ 服从正态分布$(0, \sigma)$，而非更一般的 LS 分布时，可以进行简化。对于随机选取的一个缺陷和一名检验员，潜在阈值为

$$Y = \mu + \psi + \epsilon.$$

由于 ψ 和 ϵ 是相互独立的正态随机变量，恰好能够被检测到的潜在缺陷尺寸为

$$Y \sim \mathrm{Normal}\left(\mu, \sqrt{\sigma_\tau^2 + \sigma_\delta^2 + \sigma_\gamma^2 + \sigma^2}\right).$$

因此

$$\mathrm{POD}(x) = \Pr(Y \le x) = \Phi_{\mathrm{norm}}\left[\frac{x - \mu}{\sqrt{\sigma_\tau^2 + \sigma_\delta^2 + \sigma_\gamma^2 + \sigma^2}}\right]. \tag{8}$$

3.2.8 特定检验员的条件POD

对于特定的检验员 j，检验员效应 δ_j 是固定的，而缺陷效应和缺陷/检验员交互效应是随机的。因为

$$\tau + \gamma \sim \text{Normal}\left(0, \sqrt{\sigma_\tau^2 + \sigma_\gamma^2}\right),$$

对缺陷总体进行平均后，检验员 j 的检出概率为

$$\text{POD}_j(x) = \text{E}\left[\Phi\left(\frac{x - (\mu + \delta_j + \tau + \gamma)}{\sigma}\right)\right].$$

该期望值可按照第 3.2.7 节中的方法进行数值计算。当 $\epsilon_{ijk} \sim$ 服从正态分布$(0, \sigma)$ 时，结果可简化为

$$\text{POD}_j(x) = \Phi_{\text{norm}}\left[\frac{x - (\mu + \delta_j)}{\sqrt{\sigma_\tau^2 + \sigma_\gamma^2 + \sigma^2}}\right]. \tag{9}$$

因此，正的检验员效应会使检出概率曲线向较大的缺陷尺寸方向移动，而负的检验员效应会使其向较小的缺陷尺寸方向移动。

3.2.9 分层POD模型的统计估计

虽然可以使用最大似然方法来估计本节所述随机效应模型的参数，但似然函数需要对随机效应进行复杂的高维积分计算。R 软件包（如 lme4 和 glmmTMB）通过使用似然的近似方法，避免了直接进行数值积分。计算参数的函数 （如检出概率）的估计值和近似置信界限会增加另一层复杂性，具体如第 3.2.7 节所述。此外，由于在我们的实验中，检测人员数量和重复检验次数较少，用于计算置信区间的“大样本”近似方法可能无法以所需的置信水平提供准确的近似结果。

在这些情况下，贝叶斯推断是替代最大似然估计（ML）的一种实用方法，可采用弱信息先验分布，并使用马尔可夫链蒙特卡罗（MCMC）方法计算所需后验分布的抽样。这种方法避免了 ML 软件所需的似然近似；只需在每个后验抽样处计算所需函数的值，即可轻松获得参数函数（如检出概率）的后验抽样。根据这些抽样构造的可信区间不依赖大样本正态近似理论。使用 Stan 等现代软件即可轻松采用这类方法（Stan开发团队，2026年）。Li等人Li等人（2014年）以及Koh和Meeker（2014年、2019）也曾将贝叶斯方法用于类似的检出概率应用。

用于贝叶斯分析的现代统计软件允许使用直观的建模语言来指定统计模型（包括先验分布）。对模型进行参数化并选择合适的先验分布时需要谨慎。在许多应用中，可以选择弱信息先验分布，使其在稳定弱识别参数估计的同时，允许数据中的信息占主导地位。通过适当选择先验，贝叶斯可信界限也可以具有良好的频率学派覆盖性质（例如，名义水平为 95% 的下限可信界限，其重复抽样覆盖率可约为 95%）。这些问题

已在Tian 等人（2024）中针对一些简单的可靠性应用进行了讨论和说明，但其中的思路很容易扩展到更复杂的模型。

我们使用 R 中的 Stan 接口 rstan 来执行分层POD模型的计算。根据指定的 Stan 模型和数据，Stan 会从联合后验分布中生成指定数量的抽样。对于探索性分析，当自相关不太大时，$B = 10,000$ 个抽样通常就足够了。为了准备最终结果，可能需要更多抽样，以确保报告结果中的所有有效数字都正确。对于此处给出的结果，我们使用了 $B = 40{,}000$ 个抽样。

随后使用联合后验分布的抽样来计算感兴趣量的后验抽样（例如，我们示例中的检出概率）。例如，式（7）中的检出概率是五个总体层面参数 μ, σ, σ_τ, σ_δ 和 σ_γ 的函数。设相应的后验抽样为 $\mu^{[b]}$, $\sigma^{[b]}$, $\sigma_\tau^{[b]}$, $\sigma_\delta^{[b]}$ 和$\sigma_\gamma^{[b]}$, $b = 1, \ldots, B$。这些抽样可用于计算

$$\sigma_\psi^{[b]} = \sqrt{\left(\sigma_\tau^{[b]}\right)^2 + \left(\sigma_\delta^{[b]}\right)^2 + \left(\sigma_\gamma^{[b]}\right)^2}, \quad b = 1, \ldots, B.$$

然后，对于给定的缺陷尺寸值 x，可以计算

$$\mathrm{POD}^{[b]}(x) = \int_{-\infty}^{\infty} \Phi\left[\frac{x - \left(\mu^{[b]} + \sigma_\psi^{[b]} z\right)}{\sigma^{[b]}}\right] \phi_{\mathrm{norm}}(z)\, dz, \quad b = 1, \ldots, B.$$

当$\epsilon_{ijk} \sim$ 正态分布$(0, \sigma)$ 时，可以类似地使用式(8)，从而避免数值积分。

$\mathrm{POD}^{[b]}(x)$ 的经验中位数 $b = 1$、.....、B，用作 $\mathrm{POD}(x)$ 的点估计； B 抽样的经验 0.05 分位数则是 $\mathrm{POD}(x)$ 的 95% 可信下界。对涵盖感兴趣范围的一组 x 个取值（通常 100 或 200 个已足够）重复计算，并选择相应范围，使 $\mathrm{POD}(x)$ 的可信下界从接近零延伸至接近一。然后，可将所得点估计和可信下界相对于 x 作图。通过插值求得点估计达到 0.90 时的 a_{90} 值，而通过插值求得 95% 可信下界达到 0.90 时的 $a_{90/95}$ 值。

根据第3.2.8节中的方程，也可以对特定检验员计算条件POD。例如，给定检验员 j 和缺陷尺寸 x时，式(9)计算为

$$\mathrm{POD}_j^{[b]}(x) = \Phi_{\mathrm{norm}}\left[\frac{x - (\mu^{[b]} + \delta_j^{[b]})}{\sqrt{\left(\sigma_\tau^{[b]}\right)^2 + \left(\sigma_\gamma^{[b]}\right)^2 + \left(\sigma^{[b]}\right)^2}}\right], \quad b = 1, \ldots, B.$$

4 实验研究与统计估计

本节将第3 节所述的统计方法应用于使用第2节所述流程生成的 X 射线图像所获得的检验数据。我们首先介绍检验实验及由此得到的命中-漏检数据。随后，我们使用单次检验的数据演示 POD 估计，并将结果与对完整重复测量数据的分析结果进行比较。我们使用重复测量模型估计总体均值 POD，并评估不同变异来源的相对贡献。本节

最后讨论对 XRSIM 的实验验证，以及其对模型辅助POD研究的实际意义。

4.1 实验设置

检测实验使用了通过 XRSIM 生成的合成X射线图像，这些图像用于呈现深度、宽度、位置和厚度条件各不相同的焊缝裂纹。模拟器在计算图像的同时，记录了缺陷的实际位置。文件名中嵌入了所有关键参数，并通过文件名映射到一组带有数字的文件名，该数字表示图像的查看顺序。每次查看图像时，都会重新生成该顺序。

为生成命中/未命中数据，我们使用了五名检测人员，每人将这组合成图像查看两次，两次间隔约一周。查看在同一台计算机上进行，以确保查看环境受到控制。研究中的五名检测人员中有两名缺乏经验，需要接受 MaxIm DL 查看软件的使用培训以及目标识别培训。不出所料，缺乏经验的检测人员所得结果不如经验更丰富的检测人员。对查找表（第 2.4节）的控制至关重要；在发现允许个人控制会导致漏检缺陷数量大幅增加后，我们确定需要在控制查找表变异性的条件下重新进行测试查看。早期协议开发阶段的检测结果未纳入最终数据集。

4.1.1 命中/漏检检测结果

检测记录汇总为命中/漏检结果，其中记录了存在、发现和漏检的缺陷数量。如图 8 所示，我们看到一张典型的命中/漏检图，展示了一位检验员对全部 105 个缺陷进行检测的结果。横轴变量为裂纹深度，因为这一缺陷参数决定了对比度，进而决定可探测性限值。

4.1.2 命中/漏检检测结果的定性观察

由于缺陷位于焊道的不同位置，漏检和检出的结果范围很广。在任何厚度下，紧闭的缺陷都更难发现。在较薄区域，较浅的裂纹很容易被发现；而同样的缺陷若位于较厚区域，则会因对比度不足而无法发现。值得注意的是，焊道根部的裂纹比不规则焊道上的裂纹更容易检测到。这是一种广为人知的现象：样品的不规则性会产生某种伪装效应，使最小的缺陷更难被检测到。

4.2 基于检查员1首次检查的检出概率估计

本节演示第 3.1 节中介绍的方法，其中每个缺陷仅检测一次。图 9基于图 8中所示的同一组检查员1首次检查数据（各点的垂直位置不同，是由于随机抖动结果不同），展示了基于probit二元回归模型（对应于潜在可检测缺陷尺寸变量服从正态分布）的检出概率最大似然估计值及相应的 a_{90} 值。图中还展示了95%逐点置信界限的下限及相应的$a_{90/95}$ 值。

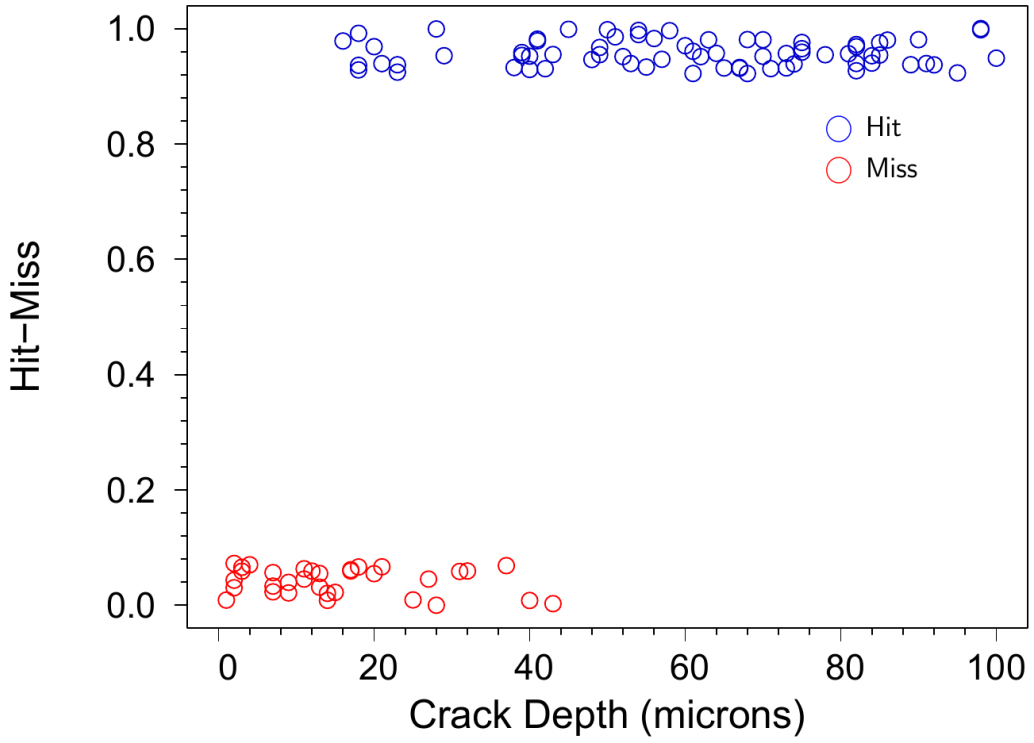


图8：检查员1对这组图像进行首次检查得到的命中-漏检数据（含抖动）。15微米至40微米的检测范围反映了可探测性的变化，这主要是由焊缝余高厚度造成的。焊缝通常不规则，其厚度变化往往与所焊接板材的厚度处于同一量级。这种变化主要是由于缺陷灵敏度对工件厚度有很强的依赖性。

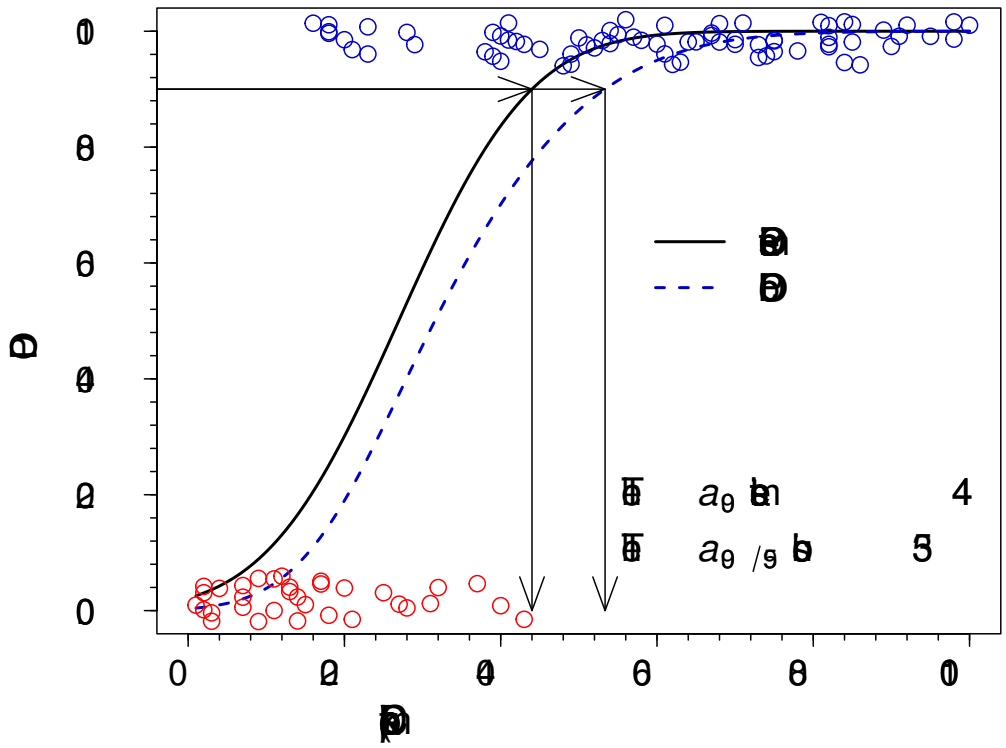


图9：基于检查员1首次检查的命中-漏检数据（含抖动），采用最大似然估计和Wald近似得到的a_{90}检出概率估计值、95%逐点下置信界（lcb）及对应的上置信界（ucb），记为$a_{90/95}$。

4.3 基于重复测量数据的POD估计

本节使用完整数据说明第 3.2 节中的统计方法。在这些完整数据中，五名检验员分别对 105 个缺陷中的每一个进行了两次检测。对部分数据进行的探索性分析（类似于第 4.2节所述的分析）表明，潜在可检测缺陷尺寸的最佳模型是正态分布，这对应于传统二元回归模型中的概率单位链接。以下本节内容均采用正态分布模型。

4.3.1 先验指定

贝叶斯统计模型要求为总体模型参数指定先验分布。我们采用了弱信息先验：在允许数据于较宽范围内主要决定后验分布的同时，将大部分概率置于适用于本应用的合理值上。这些先验分布总结于表 1中。回顾一下，所有

表 1：总体模型参数的弱信息先验总结。

参数	指定类型	分布	特征
μ	范围，自由度	位置-尺度 t	范围 $[1, 100]$；自由度 $= 3$
σ	范围，自由度	对数-位置-尺度 t	范围 $[0.1, 10]$；自由度 $= 3$
σ_τ (缺陷)	参数	半-t	尺度 $= 100$；自由度 $= 3$
σ_δ (检验员)	参数	半-t	尺度 $= 50$；自由度 $= 3$
σ_γ (缺陷/检验员)	参数	半-t	尺度 $= 10$；自由度 $= 3$

Note: “自由度”表示自由度。

总体模型参数的单位与刚好能够检测到的潜在裂纹深度水平相同（即微米），并且在这些单位下具有相应的解释。对于 μ和 σ ，我们采用Tian 等人(2024)提出的方法，并将先验参数设定为先验分布的 0.005 和 0.995 分位数所定义的 99% 范围。采用这种方法，可以方便地直接以具有实际意义的单位设定较宽的先验范围。表1 中的范围经选择后既涵盖了合理的取值，又使难以与应用情境相符的极端取值具有较低的概率。

对于随机效应标准差，我们对Gelman(2006)的建议作了修改，采用自由度为 3 的半-分布。一个通用的经验法则是：若半-分布的尺度参数是标准差点估计的十倍，那么相对于相应的边际后验分布，该先验大致是平坦的。这些先验在零附近赋予了相当大的概率，同时其相对较重的尾部也允许在数据支持时存在显著的缺陷间、检验员间或缺陷与检验员组合间变异性。我们根据先验知识选择半-分布的尺度参数：缺陷间变异性会大于其他变异来源，因此对 σ_τ采用 100 的尺度，对 σ_δ采用 50 的尺度，对σ_γ采用 10 的尺度。

4.3.2 后验分布总结

我们使用 Stan（Stan开发团队，2026）生成后验抽样。经过 10,000 次预热迭代后，我们每隔五次保留一个预热后的抽样，以降低保留抽样之间的自相关

，从而减少需要存储和处理的抽样数量，最终从联合后验分布中获得 20,000 个抽样。后验抽样的轨迹图及其他 MCMC 诊断结果（此处未显示）表明，后验抽样的混合情况良好。图 10展示了每个参数前 1,000 个后验抽样的成对图。图中显示，这些参数之间的两两后验相关性很弱。

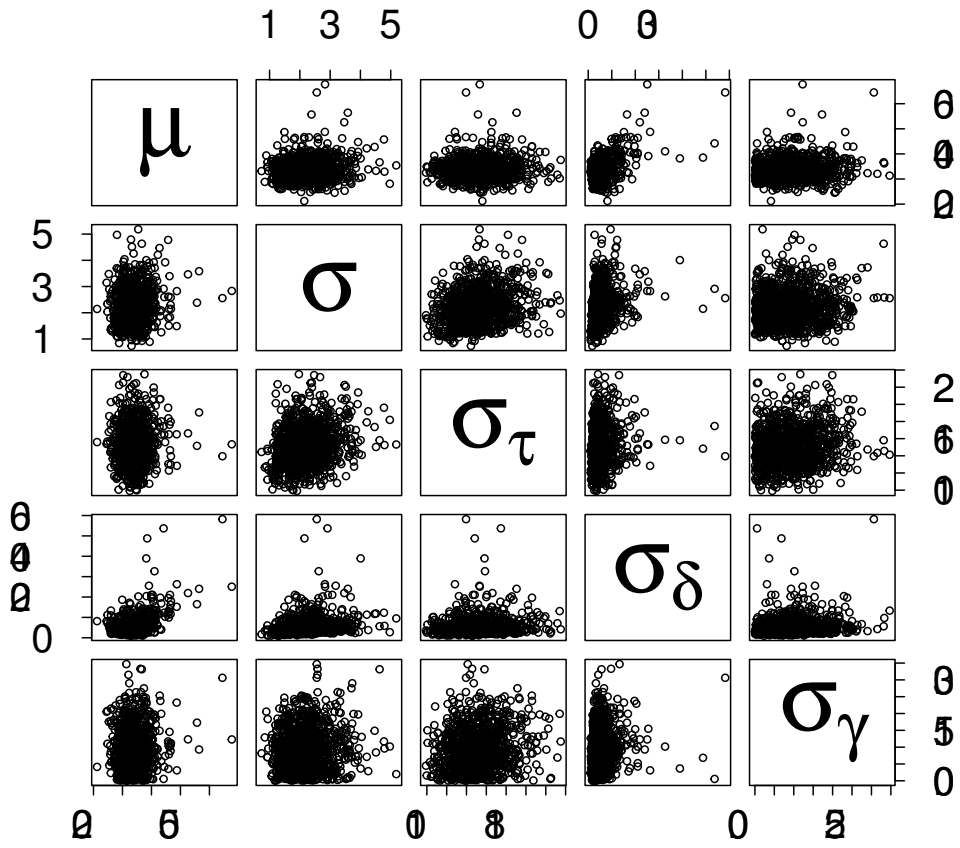


图10：P　　总体模型参数前1,000个后验抽样的配对图　　.

表2 总结了后验分布。

表2：总体模型参数的贝叶斯估计和可信区间。

参数	估计值（中位数）	95% 可信区间 下限	上限	n_{eff}
μ	33.41	26.95	42.47	30004
σ	2.09	1.17	3.57	35281
σ_τ (缺陷)	14.85	11.34	20.23	33961
σ_δ (检验员)	4.12	1.64	16.05	32367
σ_γ (缺陷检查器)	0.78	0.04	2.37	35496

$\hat{}\mu = 33.41$ 微米（临界缺陷深度的中位数）和$\hat{}\sigma = 2.09$ 微米（控制检出概率曲线上升速度）的参数描述了所有随机效应均固定为零时的基线检出概率。正如预期，$\hat{}\sigma_\tau = 14.85$ 微米（缺陷间变异性）是最主要的变异

来源。$\hat{\sigma}_{\delta} = 4.12$微米的估计值表明检验员变异性较为显著。相对较小的估计值$\hat{\sigma}_{\gamma} = 0.78$ 微米表明，与特定缺陷-检验员组合相关的额外变异性较小。由于后验抽样仍存在一些自相关， n_{eff} 的值略低于 $B = 40,000$。

4.3.3 基于重复测量数据的检出概率估计

我们使用模型参数的后验抽样来计算检出概率函数的后验抽样，从而得到检出概率估计值和可信下界，如第 3.2.9 节所述。图 11 展示了将重复测量模型拟合到数据后得到的检出概率估计值。该图

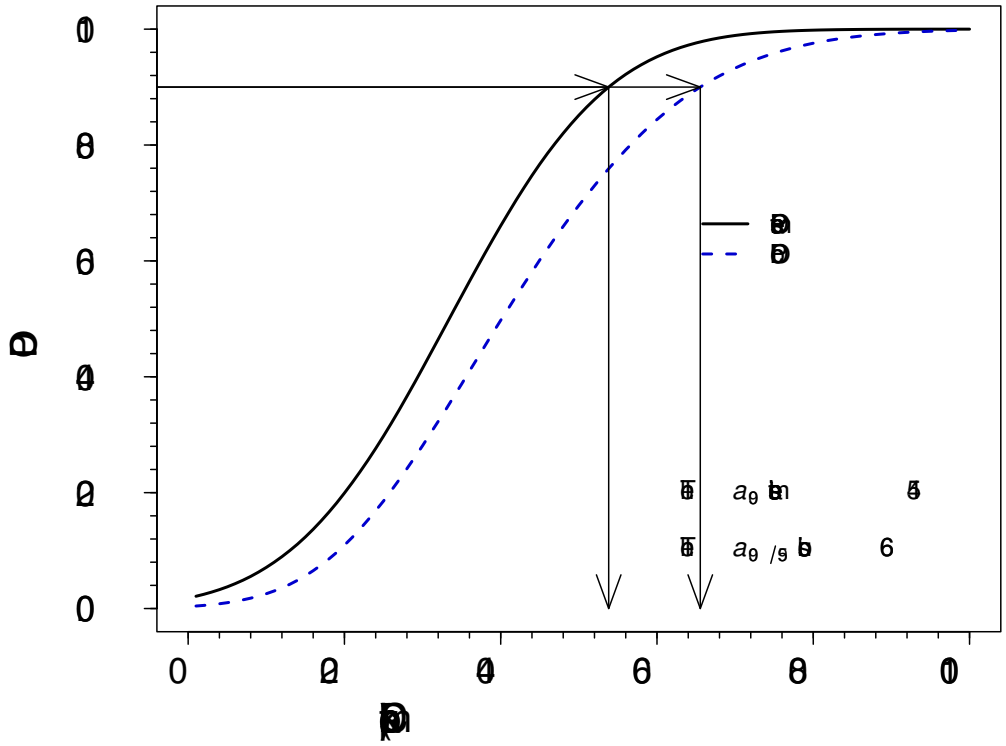


图11：基于检验数据的重复测量模型得到的总体平均检出概率估计值、95%逐点可信下界（lcb）以及对应的可信上界（ucb）；针对 a_{90}，记为 $a_{90/95}$。

还展示了 a_{90} 估计值及其对应的可信上界（ucb）$a_{90/95}$的计算过程。

4.3.4 检验员效应和检验员特定检出概率曲线的评估 es

图 12 给出了每位检验员检验员效应后验抽样的箱线图。这些图显示，五位检验员之间存在明显差异。由于这些效应表示为

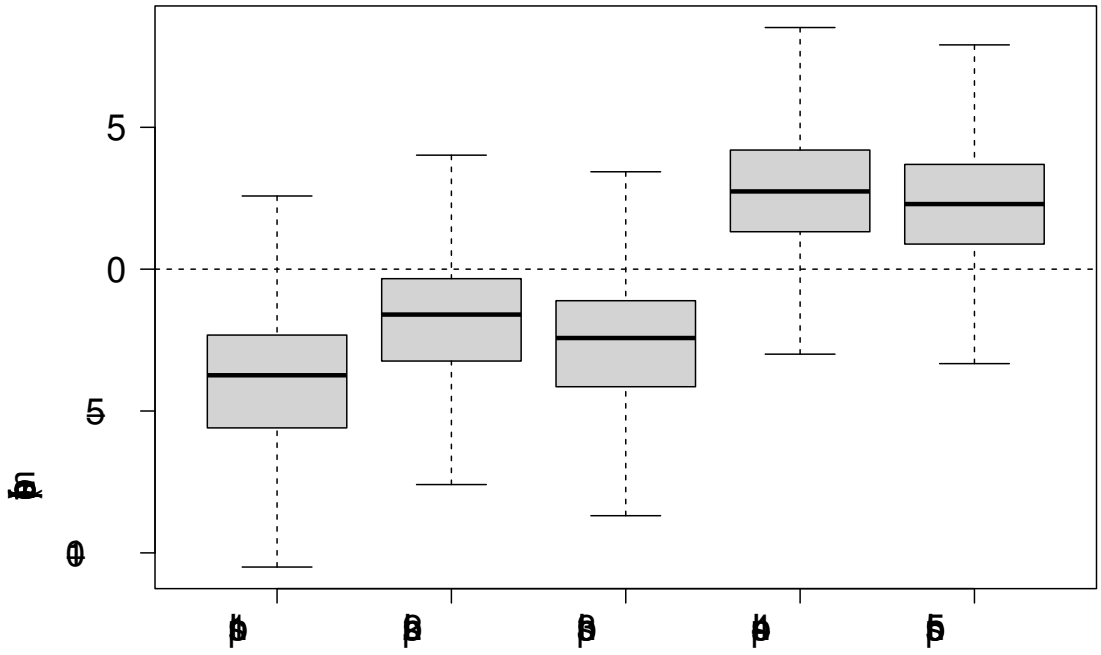


图 12：每位检验员的检验员效应后验抽样箱线图。

在潜在裂纹深度尺度上，负的检验员效应表示检测阈值低于平均水平，因此检测性能更好。正的检验员效应表示检测阈值高于平均水平，因此检测性能较差。例如，检验员 1 和检验员 4 的效应中位数之差约为 6 微米，表明检测所需裂纹深度存在具有实际重要性的检验员间变异。我们还计算了每名检验员的效应与其他检验员平均效应之间的线性对比的后验抽样。结果中（此处未显示），这些对比的 95% 可信区间对于检验员 4 和检验员 5 均不包含零，表明他们的效应不同于其他检验员的平均效应。如第 4.1 节所述，研究中的五名检验员中有两名经验不足，且检验前培训有限。

图 13a 显示了所有五名检验员的检验员特异性检出概率估计。

检验员特异性检出概率估计（针对全部五名检验员）。图 13b 显示了估计值 -

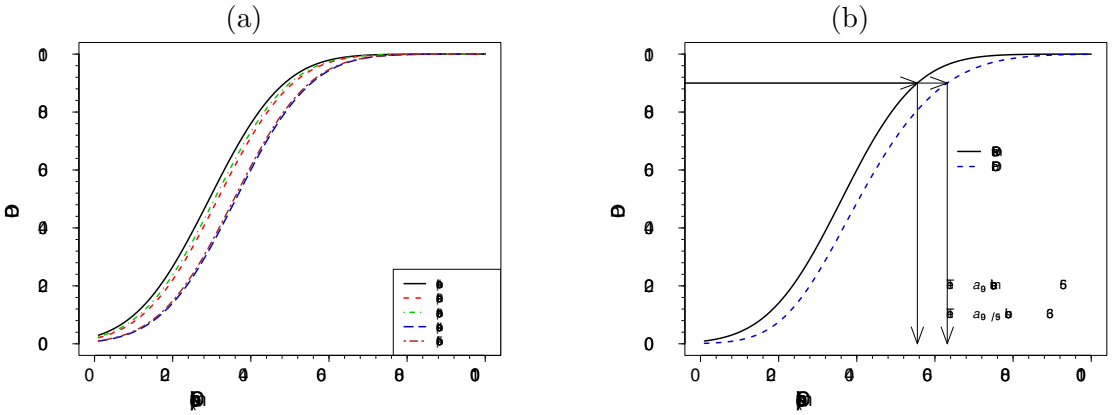


图13：全部五名检测人员的检出概率（POD）估计值（a），以及检验员 4 的 a_{90}（记为 $a_{90/95}$）的检出概率估计值、95%逐点可信下界（lcb）和相应可信上界（ucb）（b）。

检出概率和 a_{90} 的估计值，以及检验员 4 的$a_{90/95}$ 可信上界；在该实验中，检验员 4 的估计性能最差。因此，在五名检测人员中，检验员 4 给出了最保守的特定检验员检出概率曲线。

4.4 模型验证与实际意义

使用 XRSIM 的一项关键要求是对仿真模型进行实验验证。图14 和15展示了此类验证的两个示例。

图14展示了一个具有代表性的对比，反映了 XRSIM 仿真结果与实验测量结果之间观察到的一致性。

实线表示实验测量与 XRSIM 预测结果之间的吻合度在约 2% 以内。虚线表示未采用散射模型时获得的 XRSIM 结果，说明了两点：第一，实验验证的必要性；第二，不对散射建模所产生的定量影响。需要指出的是，图14所示的散射量较小。在某些涉及厚样品和高原子序数材料（如钢或镍）的检测条件下，到达探测器的 X 射线光子中，散射光子可能约占 50%。

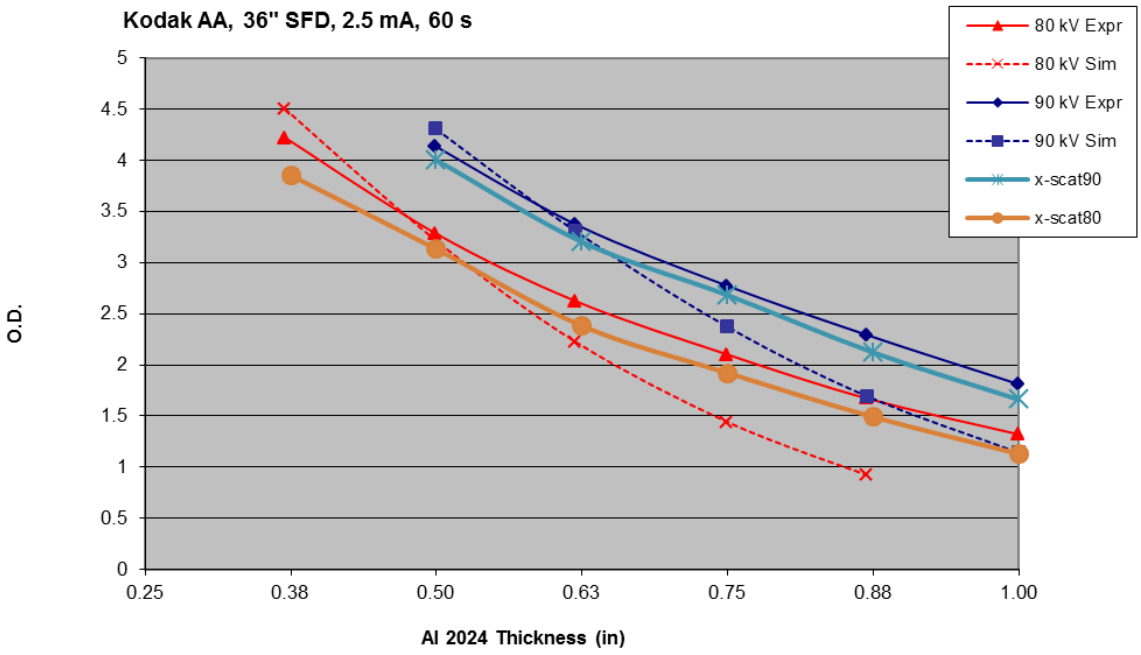


图 14：在一项工业检测验证研究中，展示了对X射线射线照相模拟程序的实验验证。实线表示在对所有关键物理变量进行建模的情况下，胶片密度随铝厚度范围变化的模拟结果和实验结果。虚线表示未使用散射模型时的模拟结果。

实验验证量化了模型预测与实验测量之间的差异。因此，对于图14所示的情况，从模型中省略散射对预测的探测器响应影响相对较小。图中未显示散射对对比度灵敏度的影响。我们指出，探测器响应的典型目标约为饱和信号的 80%。在 X射线过程中，到达探测器的 X 射线光子分为两类：直接透射的光子，其强度受工件中的衰减调制；以及对背景信号有贡献的散射光子。当探测器总响应达到饱和信号的 80% 时，直接透射部分可能只占总强度的一小部分，在极端情况下甚至不到一半。因此，尽管探测器总响应接近理想水平，承载缺陷对比度的直接透射部分却相对较弱。散射光子还会带来噪声，降低缺陷的对比噪声比（CNR）。因此，较大的散射成分会削弱检测发现较小缺陷的能力。

使用仿真模型时，重要的是了解模型对哪些物理过程进行了充分表征，并通过实验数据确定模型能够达到可接受一致性的条件范围。XRSIM能够关闭散射效应，因此在图14中生成了虚线。这一功能可快速评估散射效应的大小，为确定检测协议提供了实用工具。

第二个验证示例见图15，展示了XRSIM表征复杂缺陷形态的能力。A面板显示了从铸件气孔缺陷的三维计算机断层扫描中获得的CAD模型。计算机断层扫描数据被转换为适合导入XRSIM的三维CAD模型。对比B面板中的实验射线照片与C面板中的模拟射线照片，可以看出复杂的缺陷形态

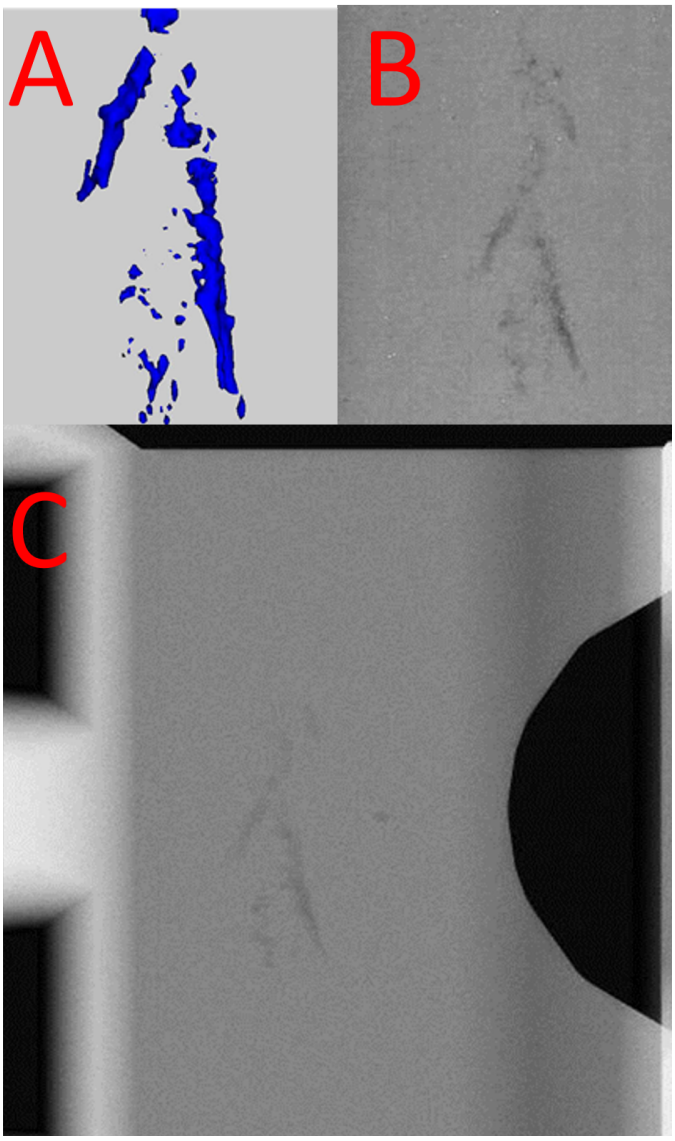


图15：展示了表示复杂且真实的缺陷几何形状的能力。A面板展示了通过对计算机断层扫描数据进行逆向工程获得的铸件气孔缺陷的计算机辅助设计（CAD）模型。B面板展示了相应的实验射线照片，C面板展示了使用 XRSIM 模拟的射线照片。

在模拟射线照片中得到了良好再现。当检出概率估计用于可靠性计算时，这种能力十分重要。很少有真实缺陷具有简单的几何形状。均匀缺口或平底孔等理想化人工缺陷有助于展示检测系统能力，但无法体现真实缺陷形态对可探测性的影响。图 15 中的结果展示了仿真工具的潜力：它们有助于区分由缺陷特征或检测条件导致的可探测性差，以及由人工检测人员造成的漏检。本研究中裂纹紧密度的重要性正说明了这一点。

5 Di讨论、结论与未来研究方向

使用 XRSIM 等无损检测仿真程序，取决于能否充分捕捉图像形成所涉及的关键物理过程。这类仿真程序必须充分体现无损检测中常用设备的输出，以及重要的工件特征，包括零件材料和典型缺陷的形态。在通过实验测定检出概率时，目标是设计一项实验，既能捕捉影响图像形成的重要变异来源，也能涵盖与检验员识别图像中特征是否为缺陷相关的人为变异性。

XRSIM能够表示检测中的许多物理变异来源。例如，可轻松纳入工件相对于X射线束的方向变化；这是检测裂纹时的一个重要因素。生成图像时，也可以改变缺陷特征和缺陷位置。探测器特性，尤其是噪声，对于确定给定缺陷的检出限至关重要。XRSIM的一项重要优势是，它能够以远比许多实验性POD研究在经济上可行的方式更详尽地模拟图像形成过程。例如，本研究所用的裂纹经过选择，旨在模拟实验性POD研究中常用人工制造缺陷较为简单的形状。然而，众所周知，这类理想化形状并不能代表老化及其他退化过程可能产生的不规则缺陷。在实验性POD研究中，制造数百个形状、方向和工件内位置范围各异的缺陷，成本将高得难以承受。使用XRSIM等工具则能更轻松地呈现此类变化。

XRSIM 本身不对人工决策过程，也不对相关的检验员间和检测间变异性进行建模。在医学成像领域，人们投入了大量精力来表征观察者表现和放射科医师变异性（例如 Fryback 和 Thornbury，1991年；Krupinski，2010年；He 和 Park，2013年）。色彩科学领域的相关工作推动了标准人类观察者模型的开发及后续完善（Berns 等人，2000年）。此外，当前分析将裂纹深度作为主要的检出概率解释变量，尽管检测响应可能取决于多种缺陷和图像特征。

因此，一个重要的机会是利用 XRSIM 生成从适当定义的检测条件总体中抽取的逼真图像，再利用这些图像直接表征人工检验员变异性。这类研究最终可提供所需信息，用于为 NDE 应用开发实用的标准人类观察者模型。同一模拟能力也为检验员培训带来重要机遇。可以让检测人员接触特定工件中可能出现的各类缺陷的逼真示例；经验丰富的检验员也可以定期复习不常见缺陷构型的相关知识。类似地，可以利用工件、缺陷和检测条件的逼真组合来培训新检验员，而不必仅仅依赖较为抽象的培训材料。

如第1.4节所述，本文提出了一种基于 XRSIM 的模型辅助POD（MAPOD）框架，用于对X射线焊缝图像进行基于图像的人工检测。本研究展示了模拟图像如何在仍能体现检测变异性重要来源的同时，减少制造大量实体缺陷试样的需求。对于重复测量检测数据，此处采用的贝叶斯随机效应模型不仅考虑缺陷间、检验员间以及缺陷/检验员组合间的变异性，还考虑检测间变异性。若将重复检查视为相互独立，就会高估数据所包含的信息量，导致统计区间过窄。本文采用的统计框架可估计总体均值POD，并为更保守的检测能力评估奠定基础，其中包括针对特定检验员的POD指标和基于分位数的POD指标。

本文的内容提出了若干值得未来研究的方向。

- Li等人(2014年) 引入了“分位数检出概率”的概念，并将其应用于涉及 $\hat{a}$ 与 a 检出概率数据的情形。 Koh和Meeker(2014年, 2019) 将类似方法应用于命中/漏检检出概率数据。其核心思想是，由于本文所述的各种变异来源，给定缺陷尺寸下的检出概率会因缺陷特征和检测条件的总体分布而异。传统POD估计会对这些变异来源取平均，从而得到总体均值POD。与产品可靠性应用中的常见做法类似（例如，Meeker等人, 2022），Li等人(2014年) 认为，在涉及安全的情况下，报告POD分布较低分位数的估计值及其置信下限可能更为恰当。此类分位数检出概率估计可通过扩展本文提出的模型来计算。

- 在许多应用中，机器学习算法会用于对图像进行初步筛查，以识别“感兴趣区域”，随后由人工检验员仔细检查这些区域并作出最终判定。本文所述的基于模拟的方法还可用于生成数据，以训练和评估机器学习算法以及检测流程中人工检验员的检测性能。

- 传统上，检出概率（POD）被估计为某一项关键缺陷特征的函数，例如裂纹长度或深度。检出概率通常是风险评估模型的重要输入，而此类模型可能需要考虑与其他缺陷特征相关的额外风险，例如方向以及缺陷在工件内的位置。原则上，POD研究可以扩展为将检出概率建模为多个解释变量的函数。使用传统实体实验开展此类研究的成本可能高得令人难以承受。对于X射线和其他基于图像的检测方法，使用图像中的模拟缺陷开展此类实验是可行的。本文开发的统计方法可以扩展用于分析此类多变量POD数据。

致谢

本文部分研究由美国联邦航空管理局通过合同编号 DTFACT-09-C-00006 资助，并在爱荷华州立大学无损评价中心开展。我们感谢 Necip Doganaksoy（锡耶纳大学）和 Luis Escobar（路易斯安那州立大学）对本文早期版本提出的宝贵意见。

利益冲突声明

作者声明不存在利益冲突。

数据可用性声明

本研究所用的数据集可在与本文相关的补充材料中获取。

参考文献

Barten, P. G. J. (1992). 人眼对比敏感度物理模型。载于*SPIE*电子成像：科学与技术研讨会，第1666卷，第57–72页。国际光学与光子学学会。 [8]

Barten, P. G. J. (1999). 人眼对比敏感度及其对图像质量的影响。华盛顿州贝灵厄姆：SPIE出版社。 [8]Berns, R. S., F. W. Billmeyer, and M. Saltzman (2000). 颜色测量。载于*Billmeyer and Saltzman*色彩技术原理，第75–105页。纽约：约翰·威立。 [28]Bismut, E. and D. Straub (2022). 一篇关于无损评价模型的综述，旨在实现最优决策支持。《结构安全》*97*，102213。 [4]

Elshafiey, A. 和 J. Gray（1996）。X射线射线照相优化工具。载于 D. Thompson 和 D. Chimemti（编），《定量无损检测进展综述》，第15A卷，纽约，第425–432页。普莱enum出版社。 [2]

Fryback, D. G. 和 J. R. Thornbury（1991年）。诊断成像的有效性。《医疗决策》*11*，88–94。 [28]Gelman, A.（2006年）。层级模型中方差参数的先验分布（对 Browne 和 Draper 文章的评论）。《贝叶斯分析》*1*，515–534。 [21]He, X. 和 S. Park（2013年）。医学成像研究中的模型观察者。《治疗诊断学》*3*，774。 [28]Jensen, T.、T. Aljundi、J. Gray 和 R. A. Wallingford（1996年）。X射线胶片响应模型。载于 D. Thompson 和 D. Chimemti（编），《定量无损检测进展综述》，第15A卷，纽约，第441–448页。普莱enum出版社。 [2]Jensen, T. 和 J. Gray（1995年）。Rtsim：实时射线照相的计算机模型。载于 D. Thompson 和 D. Chimemti（编），《定量无损检测进展综述》，第14A卷，纽约，第353页。普莱enum出版社。 [2]Kimpe, T. 和 T. Tuytschaever（2007年）。增加医学显示系统中的灰阶数量——多少才足够？《数字成像杂志》*20*（4），422–432。 [8]Koh, Y.-M. 和 W. Q. Meeker（2014年）。命中-漏检数据的分位数检出概率。载于 L. Bond 和 D. E. Chimenti（编），《定量无损检测进展综述》，第33卷，纽约州，第2023–2030页。美国物理学会。 [12, 16, 和 29]

Koh, Y.-M. 和 W. Q. Meeker（2019年）。基于命中–漏检数据的无损检测分位数检出概率。《无损检测研究》30（2），89–111。 [12, 16, 和 29]Koshti, A. M.（2014年）。用于检出概率研究的X射线过程及X射线缺陷尺寸参数建模。载于《复合材料、航空航天工程、土木基础设施与国土安全无损表征》2014，第9063卷《SPIE会议论文集》，第90631C页。[4]Koshti, A. M.（2015年）。模拟X射线图像对比度，以设置具有所需缺陷可检出性的技术。载于《先进材料、航空航天与土木基础设施结构健康监测与检测》2015，第9437卷《SPIE会议论文集》，第94370W页。 [5]Koshti, A. M.（2017年）。裂纹表面掠射角X射线裂纹检测缺陷尺寸参数建模。载于《先进材料、航空航天与土木基础设施无损表征与监测》2017，第10169卷《SPIE会议论文集》，第101691V页。[5]Krupinski, E. A.（2010年）。医学图像感知的当前观点。《注意、知觉与心理物理学》72（5），1205–1217。 [28]Le Gratiet, L.、B. Iooss、G. Blatman、T. Browne、S. Cordeiro 和 B. Goursaud （ 2017 年）。模型辅助的检出概率曲线：新的统计工具与渐进式方法。《无损检测杂志》36（1），8。 [4]Li, M.、W. Q. Meeker 和 R. B. Thompson（2014年）。用于钛锻件缺陷检测的无损检测物理模型辅助检出概率。《技术计量学》56，78–91。[4, 12, 16,和 29]Livingstone, M. 和 D. H. Hubel（2008年）。《视觉与艺术：观看的生物学》（再版）。纽约：Harry N. Abrams。 [8]Meeker, W. Q.、L. A. Escobar 和 F. G. Pascual（2022年）。《可靠性数据统计方法》（第二版）。Wiley。 [11 和 29]MIL-HDBK-1823A（2009年）。MIL-HDBK-1823A：无损检测系统可靠性评估。宾夕法尼亚州费城：国防部。 [11 和 12]Roth, D.、A. Abdul-Aziz 和 J. N. Gray（2011年）。NASA战神火箭段内蒙皮至法兰焊缝的X射线检测模拟。《材料评估》69，787–793。 [2]Stan开发团队（2026年）。《Stan用户指南》。Stan开发团队。2.39版。 [16和 21]Tian, Q.、C. Lewis-Beck、J. B. Niemi 和 W. Q. Meeker（2024年）。可靠性应用中的先验分布设定（含讨论）。《商业与工业应用随机模型》44，5–62。 [17 和 21]Xu, R.、T. Wallingford、T. Jensen 和 J. Gray（1994年）。X射线模拟代码的最新进展：XRSIM。载于 D. O. Thompson 和 D. E. Chimenti（编），《定量无损检测进展综述》，第13A卷。Plenum出版社。 [2]