Department of Geological Sciences

Backcalculation of Intelligent Compaction Data for the Mechanical Properties of Soil Geosystems

Afshin Gholamy

November 14, 2018

Outline

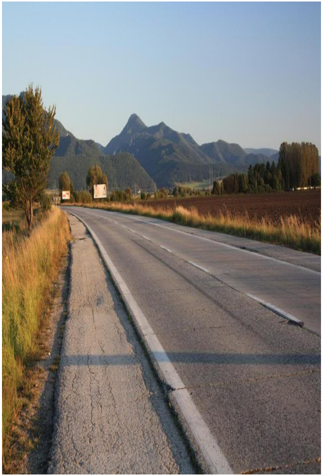
- **Formulation Of The Problem**
- **Inverse Problem for Intelligent Compaction: Results**
- **Auxiliary Tasks**
  - Elastic Modulus Formula: Theoretical Explanation
  - Safety Factors in Soil and Pavement Engineering
  - How Many Monte-Carlo Simulations Are Needed?
  - Why 70/30 Training/Testing Relation?
  - How to Minimize Relative Error?
  - How to Best Apply Neural Networks in Geosciences?
  - What Is the Optimal Bin Size of a Histogram?
- **Conclusions**

Part One

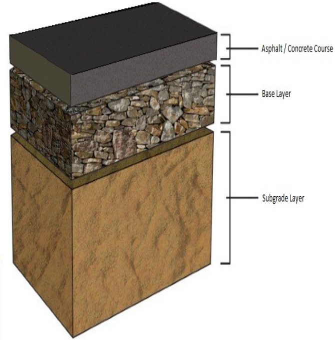
Formulation of the Problem

Need to Determine Mechanical Properties of Earthworks During Road Construction

- For national economy, it is very important to have a reliable infrastructure. So, all over the world, roads are being built, maintained, expanded, and repaired.
- Building a good quality road is very expensive, it costs several million dollars per mile. It is therefore crucial to make sure that the road lasts for a long time.
- The road is built on top of the soil.
- Soil is rarely stiff enough, so usually, the soil is first compacted.
- If needed, stiffening materials called *stabilizers* (or *treatments*) are added to the soil before compaction.



- The layer of original soil is called **subgrade.**
- On top of the subgrade, additional stiff material is placed, which is referred to as the **base** layer.
  - Base is usually composed of granular material.
  - This layer is also compacted, to make it even stiffer.
  - The base is typically reasonably thick: **15 - 30 cm**.
- It is difficult to compact a layer of such thickness, so usually, practitioners:
  - Place first a thinner layer of the base material, compact it, then
  - Place another thinner layer, etc., until they reach the desired thickness.
- Finally, **asphalt** or **concrete layer** is placed on top of the base.



Practical Problem

- For the road to be of **high quality**, all three layers must be sufficiently stiff.
- Current methods of estimating the stiffness are **time/labor-consuming**.
- A more accurate technique is:
  - Take a sample from the compacted subgrade or base, and
  - Bring it to the lab, and measure the mechanical parameters that characterize the corresponding stiffness.
- Since most roads are built in areas which are far away from the nearby labs, this procedure usually takes days.

Practical Problem (cont-d)

- While the road is being tested, the road-building company have two alternatives:
  - Keep the road building equipment idle – which will cost money, or
  - Move it to a new location, in which case there is a risk that we will need to move it back.
- To minimize this risk, companies usually **over-compact** the road – which also leads to additional costs.
- Another possibility is to measure the road stiffness **on-site**.

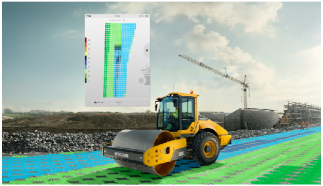
Practical Problem (cont-d)

- There are several on-site measuring techniques: LWD, FWD, DCP, NDG, PLT, etc.
- All these techniques are very labor-intensive, and take days to acquire and process the data.
- Besides, in contrast to the lab measurements:
  - These techniques do not directly measure stiffness/modulus,
  - They measure density and other parameters based on which we can only make *approximate estimates* of the desired road stiffness.

- In addition, all the existing methods – both lab-based and on-site – are **Spot tests**.
- They only gauge the road stiffness at *certain points*.
- Thus, if the road has a relatively small **weak spot**, these methods may not detect it.
- And, based on these methods, we may erroneously certify this road as ready for exploitation.
- Such a faulty road may soon require costly maintenance – at the taxpayers' expense.

Main Idea:

- We measure the road's mechanical properties while the road is being compacted by a roller.
- This can be done by placing:
  - *Accelerometers* on the rollers and/or
  - *Geophones* at different depths in several locations.
- Based on the results of these measurements, we can determine the *mechanical properties* of the road.



- The relation between the mechanical properties of the soil and the resulting accelerations is very complex.

- It is described by a **system of dynamic non-linear partial differential equations**. Even in an ideal situation:
  - When we know all the mechanical characteristics of the subgrade and of the base,
  - It takes several hours on an up-to-date computer to find the corresponding accelerations.

- We want to perform **back-calculation** (*inverse algorithm*) to determine the mechanical characteristics of the soil system from the accelerations.

- In this dissertation, we develop a method for determining the *desired characteristic in real time*.

1. **For the single-layer (subgrade) case**, we need to:
   - Determine the corresponding characteristics of stiffness based on the *acceleration measurements*.

2. **For the two-layer (subgrade + base) case**, once we have started compacting the base, we need to:
   - Determine the mechanical characteristics of the base layer based on the *measured acceleration* and on the *already-determined* characteristics of the subgrade.

Let us explain, in detail, what is needed for these tasks.

- A similar problem has been studied in the analysis of earthworks related to buildings, bridges, dams, etc., however, road-related problems are different.
  - In *building construction*, we have a reasonably *constant stress* on the underlying soil.
  - In contrast, for *road construction*, we have a *fast-changing stress* when a vehicle goes over this section of the road.
- To capture the effect of such dynamic loads, engineers developed a special notion of **elastic Modulus** (**E**).
- So, we need to estimate the elastic modulus in both layers at different locations.

- Usually, the parameters $\mathbf{k}_2'$ and $\mathbf{k}_3'$ are determined by the material – whether it is clay or gravel. In contrast, the parameter $\mathbf{k}_1'$ varies strongly even for the same material.

For example:

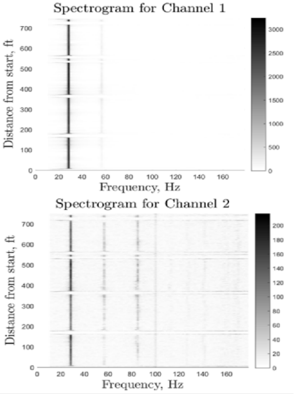
For *granular materials*, the value of $\mathbf{k}_1'$ depends on the size and shape of the grains, their density, etc. Thus:

- Once we know the substance forming the soil and/or material used for the base layer,
- We know the corresponding values $\mathbf{k}_2'$ and $\mathbf{k}_3'$ but not the corresponding values of $\mathbf{k}_1'$.

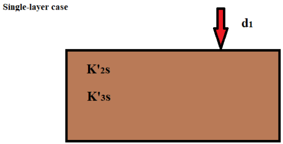
- To enhance compaction, the roller vibrates with a frequency between 20 - 40 Hz.
- So, the whole process is periodic with this frequency.

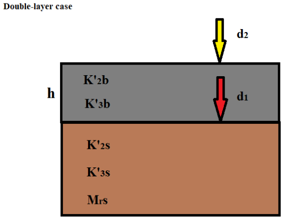
Practical Tasks

- The measured acceleration is also periodic with the same frequency.
- So, to approach this problem, it is reasonable:
  - To perform a **Fourier transform**, and to keep only the components corresponding to this known frequency.
- The resulting information can be equivalently described in terms of the **displacement (d)**.



- From this viewpoint, we face the following two tasks:
  - **Task 1** *(Single-layer case): Determine the elastic modulus* **E** *based on* $\mathbf{d_1}$, $\mathbf{k'_{2s}}$, *and* $\mathbf{k'_{3s}}$.
  - **Task 2** *(Two-layer case): Determine the elastic modulus* **E** *of the base layer from* $\mathbf{d_1}$, $\mathbf{d_2}$, $\mathbf{k'_{2b}}$, $\mathbf{k'_{3b}}$, $\mathbf{k'_{2s}}$, $\mathbf{k'_{3s}}$ *and the resilient modulus of the subgrade* $\mathbf{M_{rs}}$.





- We need *fast techniques* for solving these two problems. Therefore, we need *simple expressions* for the corresponding solutions.

- There are two ways of getting such expressions:

  1. A traditional idea is to use the corresponding **physics** to come up with possible terms.
  2. An emerging approach is to let the computers find the terms which are empirically most appropriate.

- In our research, we use both approaches and select the best of the resulting models.

First Auxiliary Task

- A challenge is that the usual mechanical equations use different mechanical characteristics:
  - Principal stresses σ_1, σ_2, and σ_3;
  - The *bulk stress* $\theta = \sigma_1 + \sigma_2 + \sigma_3$,
  - The *octahedral shear stress*

$$\tau_{\text{oct}} = \frac{1}{3}\sqrt{(\sigma_1 - \sigma_2)^2 + (\sigma_1 - \sigma_3)^2 + (\sigma_2 - \sigma_3)^2}.$$

- We thus need to know how **E** depends on $\sigma_{\mathbf{i}}$. Empirically the best model is Ooi's formula:

$$E = k_1' \left(\frac{\theta}{P_a} + 1\right)^{k_2'} \left(\frac{\tau_{\text{oct}}}{P_a} + 1\right)^{k_3'}.$$

- We want to make sure that we are not missing a more accurate expression, thus it is desirable to have a theoretical justification of Ooi's formula.

Relative Error Minimization

- Most back-calculating techniques are designed for problems where the range is small.
- In such cases:
  - To gauge the accuracy of the model, it is reasonable to simply take the difference between the actual and predicted values.

$$||\mathbf{E}(actual) - \mathbf{E}(predicted)||$$

- In pavement engineering, the elastic modulus can change by orders of magnitude.
- **For example**, the subgrade can vary between an stiff granular soil to a very soft clay.

- In such situations:
  - Large values corresponding to stiff materials will dominate, and
  - The small differences corresponding to an important case of soft subgrade will be ignored.
- It is more appropriate to use **relative errors**, i.e., approximation errors described in terms of percentages.
- In principle, we could re-write the existing software packages so that they take into account relative error.
  - This is time-consuming.
  - Some of these packages are proprietary, they do not allow to modify their code.
- It is desirable to use the **absolute-error** techniques to minimize relative errors.

Optimal Bin size of Histogram

- For engineers and practitioners to be able to use and understand the results, it is desirable to **visualize** them.
- Most of the estimates and predictions are probabilistic in nature.
- Thus, we need to plot the **histogram**, to give the user a clear understanding of the probabilities.
  - **Small bin size**, results in chaotic histogram, and does not give us a good understanding.
  - **Large bin size**, provides us with a good general picture, but we may miss important details.
- We thus need to select **optimal bin size**.

One More Auxiliary Task

- In probability and statistics, there are methods of selecting optimal bin sizes.
- However, these methods assume that we already have a *lot of information* about the probability distribution.
- In our problem, as in many other engineering tasks, we do not have this information.
- It is thus necessary to come up with a *general technique* for selecting the **optimal bin size** for a histogram.

Auxiliary Tasks: Summary

- **Task 3:** Find a theoretical justification for Ooi's empirical formula describing the resilient modulus.
- **Task 4:** Provide a theoretical justification for the empirical safety factor.
- **Task 5** Determine the appropriate number of simulations.
- **Task 6:** Come up with the most adequate division into training, testing, and validation sets.
- **Task 7:** Solve relative-error minimization problems.
- **Task 8:** Come up with neural network techniques which are most adequate for our data processing problems.
- **Task 9:** Come up with a general technique for selecting the optimal bin size for a histogram.

Part Two

Inverse Problem for Intelligent Compaction:
Main Results

Inverse Problem for Intelligent Compaction: Reminder

- Our **main objective** of this research is to develop methods for timely evaluation of the pavement quality.
- We are considering a typical case of a pavement consisting of two layers:
  - the subgrade and
  - the base (placed over this subgrade).
- First, the subgrade is reinforced (if needed) and then compacted.
- Then, the base is placed on top of the subgrade, and the pavement is compacted again.
- On each of these two compaction stages, sensors (accelerometers) are placed on the rollers.

- The accelerations measured by these sensors are then used to gauge the **quality of the pavement**.
- A good quality pavement should have:
    - a sufficiently stiff subgrade and
    - a sufficiently stiff base.
- It is important to make sure that:
    - the pavement is stiff enough, and
    - the stiffness is uniform.
- Otherwise, the traffic load will be unequally distributed.
- This will lead to too much stress (and earlier wear) for some locations.
- For the *subgrade*, its stiffness can be extracted directly from the "**pre-mapping**".

- The **stiffness of the subgrade** at each spatial location can then be determined by dividing:
    - the known force of the roller
    - by the deflection d_1 at this particular location which can be determined from the sensors attached to the compacting roller.
- The **stiffness of the base** in contrast, cannot be determined directly from the measurements.
- From the sensors attached to the roller that compacts the 2-layer pavement, we can extract the deflection d_2.

- For each spatial location, we need to evaluate the stiffness of the base from:
    - the 2-layer deflection d_2 and
    - the deflection d_1 that describes the stiffness of the subgrade.
- We can use our knowledge of materials used for the base and the subgrade layers to determine the parameters k'_2 and k'_3.
- To gauge the quality of the pavement, it is desirable to use the **“average" (representative) modulus**.
- A reasonable idea is therefore to use *the modulus at half-depth* of the base.

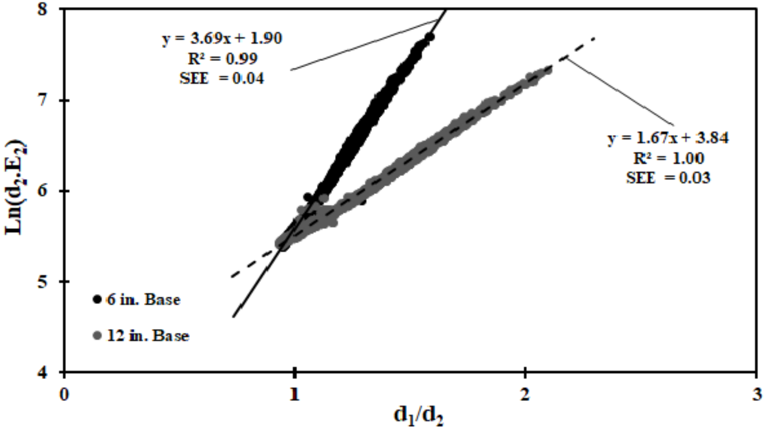
- **To test different methods** of solving the inverse problem, we used **simulations of the forward problem** for the *1-layer* and *2-layer* case.
  - We started with *linear static models* where the modulus is assumed to have the same value within each layer.
  - Next, we used *non-linear static models* in which the modulus depends on depth.
  - Finally, we used *dynamical simulations.*

- For the **1-layer case**, we only have the subgrade.
  - The Elastic modulus E_1 can be estimated based on the deflection d_1: $E_1 = \frac{c}{d_1}$, for a constant $c \approx 209$.
- In the static linear **2-layer case**, the modulus E_2 can be obtained by the following formula:

$$E_2 = \frac{1}{d_2} \cdot \exp\left(a(h) + (\ln(c) - a(h)) \cdot \frac{d_1}{d_2}\right).$$

  - In this formula, $c = 209$, and the coefficient $a(h)$ depends on the thickness h of the base:
    - for $h = 6$ inches, we have $a(h) = 1.89$;
    - for $h = 12$ inches, we have $a(h) = 3.82$;
    - for $h = 18$ inches, we have $a(h) = 4.66$.

Results of Our Analysis: Static Stationary Linear Case



- In the **static stationary nonlinear** case, we need to find the E of the base based on the following information:
  - the thichness of the base layer h
  - the displacement d_2 of the 2-layer pavement;
  - the values k'_{2b} and k'_{3b} corresponding to the base; and
  - the information about the subgrade.
- In the ideal case, we have as much information as possible about the subgrade; namely:
  - the displacement d_1 of the subgrade; and
  - the values k'_{2s} and k'_{3s} corresponding to the subgrade.
- For this case, we have come up with the following formulas.

- For the **150 mm** cases, we have

$$\begin{aligned}\ln(d_2 \cdot E) &= 2.098 + 0.361 \cdot k'_{2b} + 0.336 \cdot k'_{3b} + 0.093 \cdot k'_{2b} \cdot k'_{3b} + 0.053 \cdot (k'_{3b})^2 \\ &+ 0.467 \cdot (k'_{2s}) - 0.305 \cdot (k'_{2s})^2 - 0.264 \cdot k'_{2s} \cdot k'_{3s} - 0.079 \cdot (k'_{3s})^2 \\ &+ 0.242 \cdot k'_{2b} \cdot k'_{2s} + 0.091 \cdot k'_{2b} \cdot k'_{3s} + 0.053 \cdot k'_{3b} \cdot k'_{2s} + 3.509 \cdot \frac{d_1}{d_2} \\ &- 0.955 \cdot \left(\frac{d_1}{d_2} - 1\right)^2.\end{aligned}$$

- The R^2 is 0.95, and the mean square accuracy of this approximation is 16%.

- For the **300 mm** cases, we have

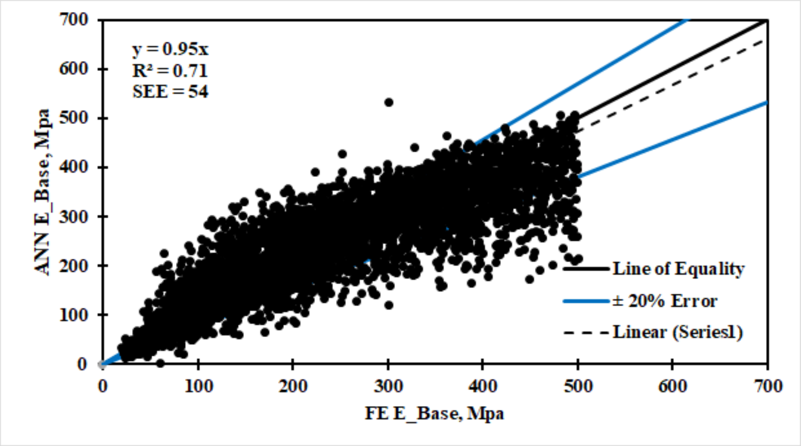
$$\begin{aligned} ln(d_2 \cdot E) &= 3.870 + 0.380 \cdot k'_{2b} + 0.348 \cdot k'_{3b} + 0.408 \cdot k'_{2s} + 0.196 \cdot k'_{3s} \\ &+ 0.078 \cdot k'_{2b} \cdot k'_{3b} + 0.037 \cdot (k'_{3b})^2 - 0.177 \cdot (k'_{2s})^2 - 0.160 \cdot k'_{2s} \cdot k'_{3s} \\ &- 0.029 \cdot (k'_{3s})^2 + 0.138 \cdot k'_{2b} \cdot k'_{2s} + 0.065 \cdot k'_{2b} \cdot k'_{3s} + 0.069 \cdot k'_{3b} \cdot k'_{2s} \\ &+ 0.041 \cdot k'_{3b} \cdot k'_{3s} + 1.656 \cdot \frac{d_1}{d_2} - 0.294 \cdot \left(\frac{d_1}{d_2} - 1\right)^2. \end{aligned}$$

- The R^2 is 0.96, and the mean square accuracy of this approximation is 11%.

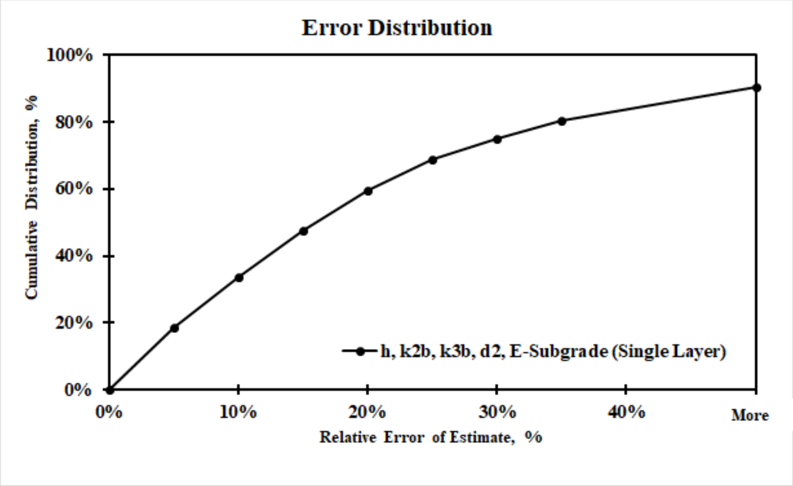
- As it was mentioned earlier, to estimate the elastic modulus E_1 of the subgrade;
    - we do not really need to know the values of the parameters k'_2 and k'_3 corresponding to the subgrade,
    - it is sufficient to know the corresponding displacement d_1.
- As a result, practitioners do not need to estimate these parameters when compacting the subgrade.
- It is therefore reasonable to also consider a realistic scenario in which:
    - instead of the values d_1, k'_{2s}, and k'_{3s},
    - we only know the estimate E_1.

- Another reason why such a scenario is needed is that:
  - while the properties of the base are reasonably well known,
  - the properties of the subgrade may vary greatly from site to site.
- In such scenario,
  - in addition to the displacement d_2, and the parameters h, k'_{2b}, and k'_{3b} that describe the base,
  - we have an estimate E_1 for the elastic modulus of the subgrade.
  - Multiple neural networks were trained to estimate the modulus E_2.
- The results are as follows:

Alternative Scenarios (cont-d)



Alternative Scenarios (cont-d)

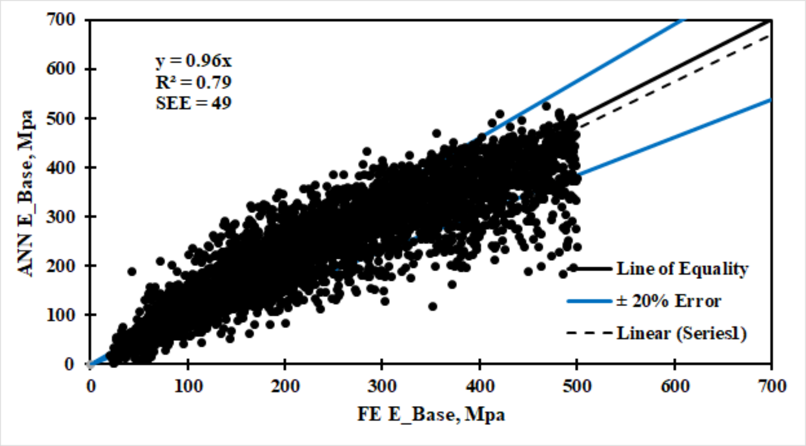


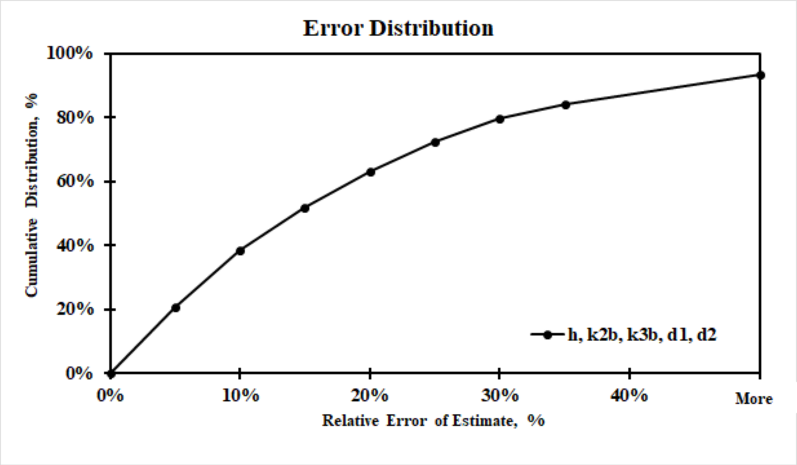
- Slightly more accurate estimates can be obtained if we take into account that the above formula for the dependence of E_1 on d_1, while reasonably accurate, is still approximate.

- Therfore, we can get better estimates if instead of using the estimate E_1, we use the displacement d_1; this way, we avoid the effect of the above inaccuracy.

- In other words, as inputs for estimating E_2, we use d_2, h, k'_{2b}, k'_{3b}, and the displacement d_1.

- The resulting estimates of training the corresponding neural network model are presented here:


y = 0.96x
R² = 0.79
SEE = 49
ANN E_Base, Mpa
FE E_Base, Mpa
0
100
200
300
400
500
600
700
Line of Equality
± 20% Error
Linear (Series1)


Error Distribution
100%
80%
60%
40%
20%
0%
Cumulative Distribution, %
0%
10%
20%
30%
40%
More
Relative Error of Estimate, %
h, k2b, k3b, d1, d2

- Even more accurate estimates for E_2 can be obtained if we use the actual values E_1^{act} of the subgrade's representative modulus E_1 instead of d_1 or the d_1-based estimate for E_1.

- In this case, to estimate E_2, we use the values d_2, h, k'_{2b}, k'_{3b}, and E_1^{act}.

- The results of the corresponding neural network model are presented here:

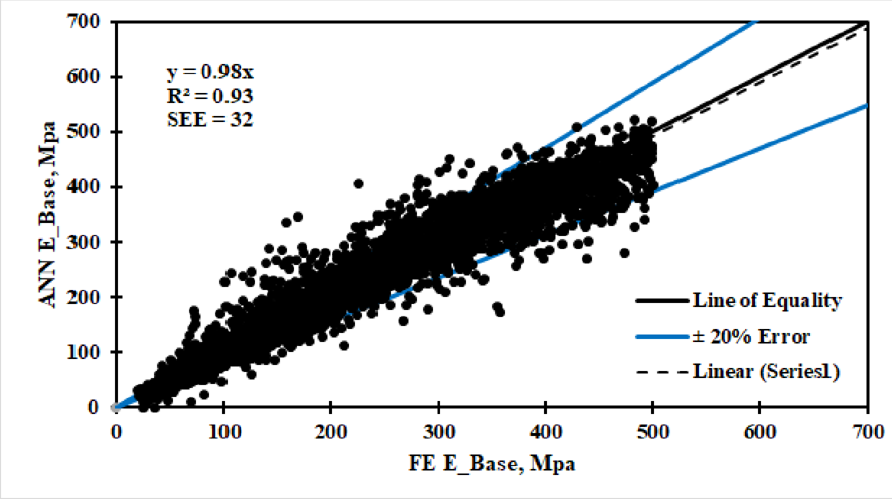


Figure: Case when we use the actual value E_1^{act}

Alternative Scenarios

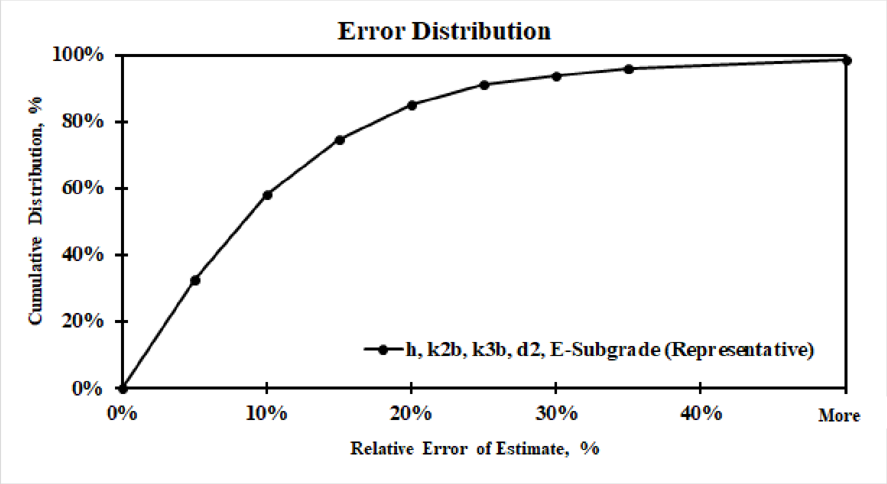


Figure: Case when we use the actual value E_1^{act}

Alternative Scenarios

- We combined the distribution of the estimation errors of all three models in a single graph.

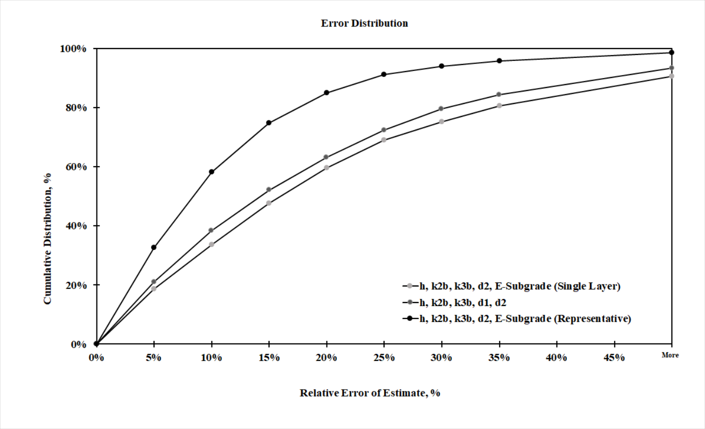


Figure: Comparative accuracy of three neural network models

- All the simulations assume that both the subgrade and the base layers are homogeneous.
- In practice, the mechanical properties of the subgrade and of the base can randomly change by 20-25%, so it is sufficient to have accuracy 20-25%.
- It turns out that with this accuracy, we can have

$$\ln(d_2^{\text{dyn}} \cdot E^{\text{dyn}}) = a_0 + a_1 \cdot \ln(d_2^{\text{stat}} \cdot E^{\text{stat}}).$$

  - For **150 mm**, $a_0 = 0.96$ and $a_1 = 0.81$, accuracy is 17%.
  - For **300 mm**, $a_0 = 1.68$, $a_1 = 0.69$, accuracy 16%.
- Thus, time-consuming dynamical simulations are not needed, and we can reconstruct their results based on static cases.

Part Three

Theoretical Explanation of Ooi's Formula

- Experimental comparison shows that the best description is provided by the formula

$$E = k_1' \cdot \left(\frac{\theta}{P_a} + 1\right)^{k_2'} \cdot \left(\frac{\tau_{\text{oct}}}{P_a} + 1\right)^{k_3'},$$

where

$$\theta = \sigma_1 + \sigma_2 + \sigma_3$$

and

$$\tau_{\text{oct}} = \frac{1}{3} \cdot \sqrt{(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2}.$$

- How can we explain this formula?

Main Idea

- Computers process numerical values of different quantities.
- A numerical value of a quantity depends:
  - on the choice of a **measuring unit**, and
  - on the choice of **the starting point**.
- For example: we can describe the height of the same person as 1.7 m or 170 cm.
  - 14.00 by El Paso time is 15.00 by Austin time.
  - Reason: the starting points – midnights (00.00) – differ by an hour.
  - The choice of a measuring unit is rather arbitrary.

Main Idea (cont-d)

- It is reasonable to require that the fundamental physical formulas not depend
  - on the choice of a measuring unit and
  - if appropriate – on the choice of the starting point.
- We do not expect that, e.g., Newton's laws look differently if we use meters or feet.
- If we change the units, then we may need to adjust units of related quantities.
  - For example, if we replace m with cm, then we need to replace m/sec with cm/sec when measuring velocity.
- However, once the appropriate adjustments are made, we expect the formulas to remain the same.

Not All Physical Quantities Allow Both Changes

- Some quantities have a fixed starting point.
- Examples:
  - while we can choose an arbitrary starting point for time,
  - for distance, 0 distance seems to be a reasonable starting point.
- As a result:
  - while the change of a measuring unit makes sense for most physical quantities,
  - the change of a starting point only makes sense for some of them.
- A physics-based analysis is needed to decide whether this change makes physical sense.

- **Case:** we replace the original measuring unit with a new unit which is a times smaller.
- Then all numerical values of the measured quantity get multiplied by a: $x' = a \cdot x$.
  - *Example:* 1.7 m is $x' = a \cdot x = 100 \cdot 1.7 = 170$ cm.
- **Case:** we replace the original starting point by a new one which is b earlier (or smaller).
- Then to all numerical values of the measured quantity the value b is added: $x' = x + b$.
- 14 hr in El Paso is $x' = x + b = 14 + 1 = 15$ in Austin.
- **General case:** we can change both the measuring unit and the starting point. Then, $x \to a \cdot x + b$.

- **For E**, there is a clear starting point $E = 0$, in which strain does not cause any stress.

- So, for E, only a change in a measuring unit makes physical sense.

- **For θ**, a change in starting point is also possible:
  - we can only count the external stress,
  - or we can explicitly take atmospheric pressure into account.

- It turns out that Ooi's formula can be derived from these invariances.

Part Four

Safety Factors in Soil and Pavement Engineering: Theoretical Explanation of Empirical Data

What Is a Safety Factor

- Models are approximations to reality.
- To describe a complex real-life process by a feasible model:
  - we find the most important factors affecting the process and
  - we model them.
- The ignored factors are smaller than the factors that we take into account; however:
  - they still need to be taken into account
  - if we want to provide guaranteed bounds for the desired quantities.
- To take these small factors into account, engineers multiply the results of the model by a constant.
- This constant is known as the **safety factor**.

- In many applications, a safety factor is 2 or smaller.
- However, in soil and pavement engineering, the situation is different.
- Researchers compared:
  - the elastic modulus predicted by the corresponding model and
  - the modulus measured by Light Weight Deflectometer.
- This comparison showed that:
  - to provide guaranteed bounds,
  - we need a safety factor of 4.
- How can we explain this?

- Let Δ be the model's estimate.
- When designing the model, we did not take into account some factors. Let's denote the effect of the largest of these factors by Δ_1.
- The factors that we ignored are smaller than the one we took into account, so:

$$\Delta_1 < \Delta, \text{ i.e., } \Delta_1 \in (0, \Delta).$$

- We do not have any reason to assume that any value from the interval $(0, \Delta)$ is more frequent than others. Thus, it makes sense to assume that Δ_1 is uniformly distributed on $(0, \Delta)$.
- Then, the average value of Δ_1 is $\dfrac{\Delta}{2}$.
- The next smallest factor Δ_2 is smaller than Δ_1.

- The same arguments shows that its average value of Δ_2 is $\dfrac{\Delta_1}{2}$, i.e.,

$$\begin{aligned} \Delta_1 &= 2^{-1} \cdot \Delta \\ \Delta_2 &= 2^{-1} \cdot \Delta_1 \\ &= 2^{-2} \cdot \Delta \end{aligned}$$

Therefore, for each $k > 0$

$$\Delta_k = 2^{-k} \cdot \Delta$$

- Hence the overall estimate is

$$\begin{aligned} \Delta + \Delta_1 + \ldots + \Delta_k + \ldots &= \Delta + 2^{-1} \cdot \Delta + \ldots + 2^{-k} \cdot \Delta + \ldots \\ &= 2\Delta. \end{aligned}$$

- Empirical data shows that for soil and pavement engineering, 2 is not enough.
- This means that Δ_1 should be larger than our estimate $\frac{\Delta_1}{2}$: $\Delta_1 \in (\frac{\Delta}{2}, \Delta)$.

$$\begin{aligned}\Delta_1 &= \frac{3}{4} \cdot \Delta \\ \Delta_2 &= \left(\frac{3}{4}\right)^2 \cdot \Delta\end{aligned}$$

Therefore, for each $k > 0$

$$\Delta_k = \left(\frac{3}{4}\right)^k \cdot \Delta$$

and thus,

$$\begin{aligned}\Delta + \Delta_1 + \ldots + \Delta_k + \ldots &= \Delta \cdot (1 + 3/4 + \ldots + (3/4)^k + \ldots) \\ &= \Delta/(1 - 3/4) \\ &= 4\Delta.\end{aligned}$$

Part Five

How Many Monte-Carlo Simulations Are Needed?

- We provide a partial answer to the question of how many Monte-Carlo simulations are needed.
- Namely, we provide this answer for the case of interval uncertainty.
- A recent research used Monte-Carlo technique to compare algorithms for smart electric grid.
- The study computed ranges (intervals) for different quantities.
- This study shows that we need ≈ 500 simulations to reach 5% accuracy.
- In this dissertation, we provide a theoretical explanation for these empirical results.

Part Six

Why 70/30 Relation Between Training and Testing Sets

Why 70/30 Relation Between Training and Testing Sets?

- Overfitting is when the model closely follows all the available data – but is lousy in predictions.
- To avoid this, it is important to divide the data into the **training set** and the **testing set**.
  - We first train our model on the training set.
  - Then we use the data from the testing set to gauge the accuracy of the resulting model.
- Empirical studies show that the best results are obtained if we use **20-30%** of the data for testing. The remaining **70-80%** of the data is for training.
- We provide a possible explanation for this empirical result.

Part Seven

How to Minimize Relative Error

How to Minimize Relative Error?

- In many problems, there is a need to find the parameters of a dependence from the experimental data.
- There exist several software packages that find the values for these parameters.
- Usually, we minimize the the **mean square value** of the **absolute approximation error**.
- For example, in the linear case, we minimize the sum

$$\sum_{k=1}^{K} \left(y^{(k)} - \sum_{j=1}^{m} c_j \cdot f_j \left(x_1^{(k)}, \ldots, x_n^{(k)} \right) \right)^2 .$$

- In practice, however, we are often interested in minimizing the **relative approximation error**.

- It is desirable to use absolute-value software to minimize relative errors.
- We analyze this problem.
- For the **linear case**, our recommendation is to minimize the sum

$$\sum_{k=1}^{K}\left(1-\sum_{j=1}^{m} c_j \cdot \frac{f_j\left(x_1^{(k)},\ldots,x_n^{(k)}\right)}{y^{(k)}}\right)^2.$$

- A similar recommendation works in the **non-linear case.**

Part Eight

How to Best Apply Neural Networks

How to Best Apply Neural Networks

- The main objectives of geosciences is:
  - to find the current state of the Earth – i.e., solve the corresponding *inverse problems* – and
  - to use this knowledge for predicting the future events, such as earthquakes and volcanic eruptions.
- In both inverse and prediction problems, often, machine learning techniques are very efficient.
- At present, the most efficient machine learning technique is **neural networks**.
- To speed up this training, the current machine learning algorithms use **dropout techniques**:
  - they train several sub-networks on different portions of data, and
  - then "average" the results.

- A natural idea is to use **arithmetic mean** for this "averaging".
- However, empirically, **geometric mean** works much better.
- In this dissertation, we explain the empirical efficiency of geometric mean.
- This explanation uses the same independence-on-measuring units as in explaining Ooi's formula.

Part Nine

What Is the Optimal Bin Size of a Histogram?

What Is the Optimal Bin Size of a Histogram?

- A natural way to estimate the probability density function from the sample is to use **histograms**.
- The accuracy of the estimate depends on the size of the histogram's bins h.
- There exist heuristic rules for selecting the bin size.
- The probability density $\rho(x)$ changes from 0 to 1 then from 1 to 0 on an interval of width s. Therefore, its average rate of change is $1/(s/2)$.
- In a bin of size h, we approximate each value $\rho(x)$ by a value at midpoint.
- The largest distance from midpoint is $h/2$.

- Thus, the relative change is $(h/2)/(s/2) = \frac{h}{s}$.
- On the other hand, in each bin, we have $m = n \cdot \frac{h}{s}$ points.
- According to statistics, estimates based on m points has accuracy $\frac{1}{\sqrt{m}}$.
- Minimizing the overall error $\frac{h}{s} + \frac{1}{\sqrt{m}}$
- Therefore:

  $$h_{\text{opt}} = \text{const} \cdot s \cdot n^{-1/3}$$

  which is exactly the empirically **optimal bin size**.
- Thus, we have theoretically justified this empirical choice.

Part Ten

Conclusions

Conclusion

- It is desirable to speed up the quality assessment of the newly built roads.Thus, we need to measure the road's stiffness in real time, as the road is being built.

- This is the main idea behind **intelligent compaction**, when:
  - accelerometers and other measuring instruments are attached to the roller, and
  - the results of the corresponding measurements are used to gauge the road's stiffness.

- **The main challenge:** formulas for estimating the road's stiffness are complicated.

- It is a complex system of partial differential equations which are difficult to solve in real time.

Conclusion

- It is therefore desirable to come up with easier-to-compute algorithms.
- The **main task** of this dissertation was the design of such algorithms.
- As a solution to this task, we propose both:
  - most-easy-to-compute analytical expressions and
  - somewhat more complex (but still easy to compute) neural network models.
- We also provide a theoretical explanations for:
  - the empirical formulas used to describe the road dynamics, and
  - the empirical safety factor that related the simulation results with actual road measurements.

Conclusion

- In the process of solving the main task, we have also solved several auxiliary tasks, explaining:

  - how many simulations are needed,
  - what is the best relation between training and testing sets,
  - how to take into account that we often need to minimize relative error,
  - how to best apply neural networks, and
  - what is the optimal bin size in a histogram.

- We hope that both our solution of the main tasks and our solutions to the auxiliary tasks will be useful.

Acknowledgments

- I am very thankful to Drs. Laura Serpa and Vladik Kreinovich my Committee Co-Chair.
- I am also very grateful to all the members of my committee:
  - Soheil Nazarian,
  - Hector Gonzalez,
  - and last but not least, Aaron Velasco.
- Many thanks to all for your help and support.