Why agricultural and environmental research still need statisticians

R. A. Bailey

University of St Andrews

QMUL (emerita)

Statistical challenges and opportunities for modern agricultural and environmental research
Royal Statistical Society International Conference 2023
Harrogate
7 September 2023

Overview

When I worked at Rothamsted in the 1980s, statisticians were involved in all stages of research, from planning the experiment to analysing the data and drawing conclusions. An experiment was not given permission to go ahead unless it had been approved by a statistician. The Statistics Department would not give their approval unless one of the statisticians had done a dummy analyis on dummy data to make sure that there were no unforeseen confoundings in the proposed design.

Overview

When I worked at Rothamsted in the 1980s, statisticians were involved in all stages of research, from planning the experiment to analysing the data and drawing conclusions. An experiment was not given permission to go ahead unless it had been approved by a statistician. The Statistics Department would not give their approval unless one of the statisticians had done a dummy analyis on dummy data to make sure that there were no unforeseen confoundings in the proposed design.

In the 1990s, suddenly there was a common mantra that "We have all got computers, so who needs statisticians?" Several research organizations made all of their statisticians redundant.

Overview

When I worked at Rothamsted in the 1980s, statisticians were involved in all stages of research, from planning the experiment to analysing the data and drawing conclusions. An experiment was not given permission to go ahead unless it had been approved by a statistician. The Statistics Department would not give their approval unless one of the statisticians had done a dummy analyis on dummy data to make sure that there were no unforeseen confoundings in the proposed design.

In the 1990s, suddenly there was a common mantra that "We have all got computers, so who needs statisticians?" Several research organizations made all of their statisticians redundant.

I will give two examples, one from ecology and one from agriculture, to show that statisticians are still needed.

First problem: false replication

Sometimes the experimenter applies each treatment to a single large unit, and then takes several measurements within that unit. Then there is no way of distinguishing between differences between treatments and differences between large units.

First problem: false replication

Sometimes the experimenter applies each treatment to a single large unit, and then takes several measurements within that unit. Then there is no way of distinguishing between differences between treatments and differences between large units.

This is called false replication. The problem has been pointed out many times, but it still goes on.

Example 1: Ecology

In August 2017 I receive an email from an Ecology PhD student in California. She tells me that she has done an experiment measuring the consumption of kelp by various combinations of beach detrivores.
She has seen two of my papers about modelling biodiversity. Can I advise her how to analyse her data? She attaches some ANOVA summaries giving values of F and p only.

Example 1: Ecology

In August 2017 I receive an email from an Ecology PhD student in California. She tells me that she has done an experiment measuring the consumption of kelp by various combinations of beach detrivores.
She has seen two of my papers about modelling biodiversity.
Can I advise her how to analyse her data? She attaches some ANOVA summaries giving values of F and p only.

I ask her to show me the complete ANOVA tables.
I also advise her to fit the family of models that I used in my previous papers.

Example 1: Ecology

In August 2017 I receive an email from an Ecology PhD student in California. She tells me that she has done an experiment measuring the consumption of kelp by various combinations of beach detrivores.
She has seen two of my papers about modelling biodiversity.
Can I advise her how to analyse her data? She attaches some ANOVA summaries giving values of F and p only.

I ask her to show me the complete ANOVA tables.
I also advise her to fit the family of models that I used in my previous papers.

In September she sends me the "complete" ANOVA tables, but these are one table per model, not a family of nested models.
She would like me to tell her how to run my analysis in R.

Example 1: Ecology

In August 2017 I receive an email from an Ecology PhD student in California. She tells me that she has done an experiment measuring the consumption of kelp by various combinations of beach detrivores.
She has seen two of my papers about modelling biodiversity.
Can I advise her how to analyse her data? She attaches some ANOVA summaries giving values of F and p only.

I ask her to show me the complete ANOVA tables.
I also advise her to fit the family of models that I used in my previous papers.

In September she sends me the "complete" ANOVA tables, but these are one table per model, not a family of nested models.
She would like me to tell her how to run my analysis in R.

I try to tell her how to do so. This is rather brief, as I am travelling.

My summary of the treatments

The treatments consisted of 57 different combinations of 6 species, with 12 individuals in each combination.
So there were

6	monocultures, each with 12 of a single species
15	duocultures, each with $6 + 6$ of two species
20	tricultures, each with $4 + 4 + 4$
15	quadricultures (I made up this word), each with $3 + 3 + 3 + 3$
1	sextoculture (ditto) with $2 + 2 + 2 + 2 + 2 + 2$
57	in total

Each treatment was replicated 3 times.

My suggested family of models

For each treatment, let x_i denote the number of species i present, for $i = 1, \ldots, 6$. Thus $x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 12$. Let j denote the number of different species present in the treatment. This number is called the level of richness.

Constant	There is a constant c such that $y = c$ for all treatments.
Type	There are constants a_1, a_2, a_3, a_4, a_5 and a_6 such that $y = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5 + a_6x_6$ no matter the level of richness.
Richness	There are constants r_1, r_2, r_3, r_4 and r_6 such that $y = r_j$ for every treatment with richness level j.
Richness+Type	$y = r_j + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5 + a_6x_6$.
Richness∗Type	There are constants a_{ij} for $1 \leq i \leq 6$ and every level j of richness such that, for every treatment with level j of richness, $y = a_{1j}x_1 + a_{2j}x_2 + a_{3j}x_3 + a_{4j}x_4 + a_{5j}x_5 + a_{6j}x_6$.
Combination	Each of the 57 treatments gives a different expectation.

More emails

After a month away on a field trip, she sends me her anova output. It is clear that she is not fitting my suggested models.

More emails

After a month away on a field trip, she sends me her anova output. It is clear that she is not fitting my suggested models.

In October I explain the family of models in more detail. I also ask

> *Did you completely randomize the positions of the* 171 *objects, or did you separate them into three blocks, with one replicate of each treatment per block?*

More emails

After a month away on a field trip, she sends me her anova output. It is clear that she is not fitting my suggested models.

In October I explain the family of models in more detail. I also ask

> *Did you completely randomize the positions of the* 171 *objects, or did you separate them into three blocks, with one replicate of each treatment per block?*

She responds

> *There was no block design, all* 3 *reps for each treatment were run at the same time, not during different rounds/blocks.*

More emails

After a month away on a field trip, she sends me her anova output. It is clear that she is not fitting my suggested models.

In October I explain the family of models in more detail. I also ask

> *Did you completely randomize the positions of the* 171 *objects, or did you separate them into three blocks, with one replicate of each treatment per block?*

She responds

> *There was no block design, all* 3 *reps for each treatment were run at the same time, not during different rounds/blocks.*

I ask

> *Does that mean that you ran all* 171 *combinations at the same time? Or did you do only one treatment at a time? Or something in between?*

Weeks

She reponds

> *I ran the 1 and 2 species treatments first, the 3 species treatments the following week and the 4 and 6 species treatments the third week.*

Weeks

She reponds

> *I ran the 1 and 2 species treatments first, the 3 species treatments the following week and the 4 and 6 species treatments the third week.*

I explain that this means that there is no way to know if any difference between tricultures and the others is caused by a difference between weeks.

Weeks

She reponds

> *I ran the 1 and 2 species treatments first, the 3 species treatments the following week and the 4 and 6 species treatments the third week.*

I explain that this means that there is no way to know if any difference between tricultures and the others is caused by a difference between weeks.

She tries to argue that they did this all in the lab, which does not change from week to week.

Weeks

She reponds

> *I ran the 1 and 2 species treatments first, the 3 species treatments the following week and the 4 and 6 species treatments the third week.*

I explain that this means that there is no way to know if any difference between tricultures and the others is caused by a difference between weeks.

She tries to argue that they did this all in the lab, which does not change from week to week.

I persist in asking questions.

She tells me that species were collected from the beach on the first morning of each week. Her team put kelp and sand into 57 plastic tubs, then put one species combination into each. After 72 hours they removed all the detrivores and measured the weight of the remaining kelp.

So we need to consider weeks

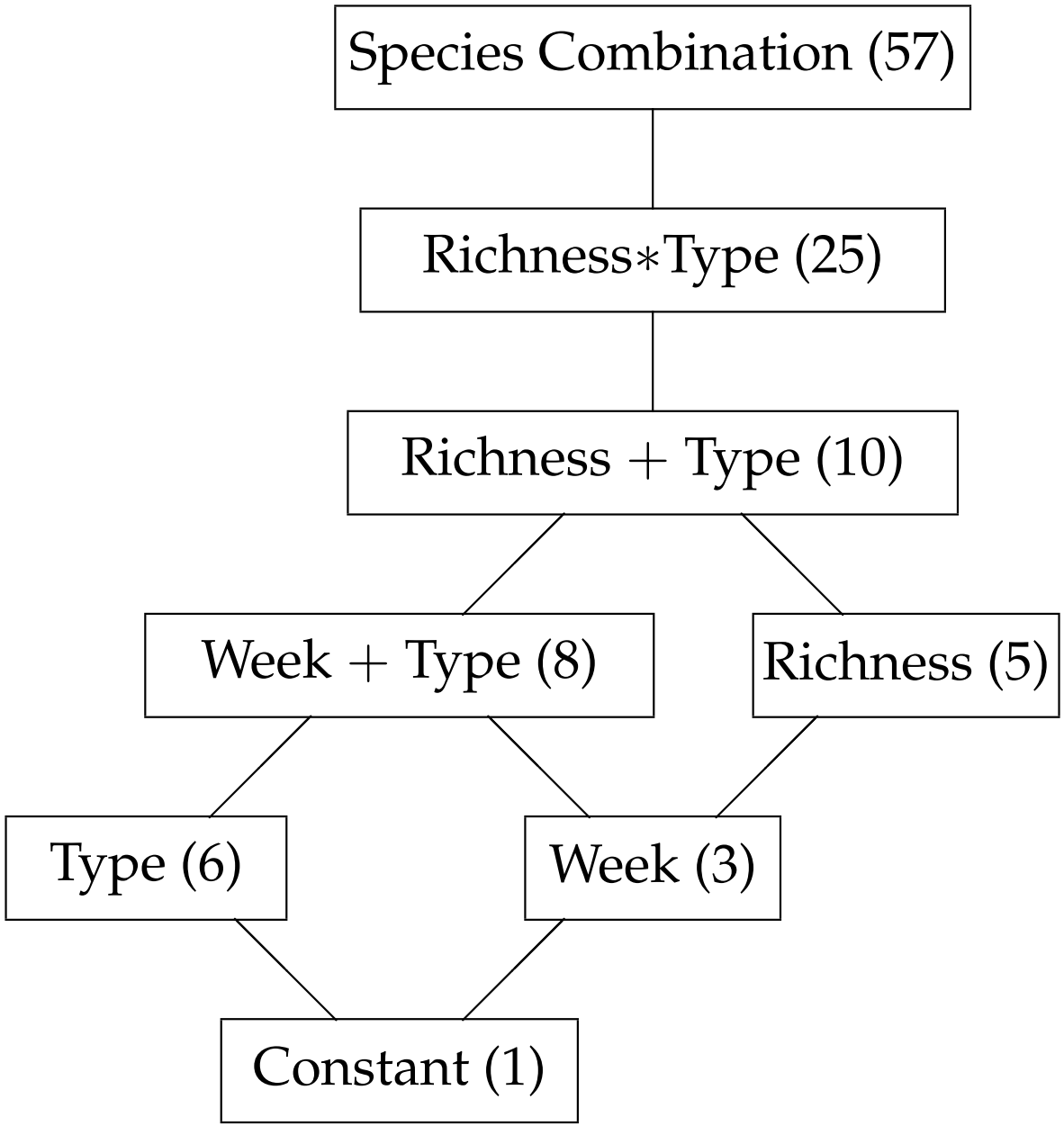
So, all the treatments in one week are collected at the same time;
they are essentially dealt with a single batch;
and finally I am told that different undergraduates helped with the lab work each week.

So we need to consider weeks

So, all the treatments in one week are collected at the same time;
they are essentially dealt with a single batch;
and finally I am told that different undergraduates helped with the lab work each week.

So I need to add Weeks to the family of models.

Hierarchy of models (numbers show dimensions)



Subsequent data analysis

Perhaps fortunately,
the F-ratio for the difference between the models

$$\text{Species Combination and Richness} * \text{ Type}$$

was 3.18 on 32 and 114 degrees of freedom.

Subsequent data analysis

Perhaps fortunately,
the F-ratio for the difference between the models

$$\text{Species Combination and Richness} * \text{Type}$$

was 3.18 on 32 and 114 degrees of freedom.

So there was no way that we could simplify the model,
so we did not have to argue about, or explain, the role of Weeks.

Second problem: computer search or expert knowledge

Expert knowledge: we all know that equal replication of treatments is best.
Computer search: this is not true for an experiment on v treatments with v blocks of size 2 when $v \geq 9$.

Second problem: computer search or expert knowledge

Expert knowledge: we all know that equal replication of treatments is best.
Computer search: this is not true for an experiment on v treatments with v blocks of size 2 when $v \geq 9$.

So should we always use computer search?
Or should we use expert knowledge about combinatorical patterns and theorems about what is best?

Example 2: A two-phase experiment in agriculture

The treatments are 10 varieties of common beans.

Example 2: A two-phase experiment in agriculture

The treatments are 10 varieties of common beans.

In Phase I, these are grown in a field, in 10 blocks of size 6.

Example 2: A two-phase experiment in agriculture

The treatments are 10 varieties of common beans.

In Phase I, these are grown in a field, in 10 blocks of size 6.

In Phase II, a sample of beans is taken from each plot.
Each sample is cooked in a special machine. The measured response is the time taken to properly cook the beans.

Example 2: A two-phase experiment in agriculture

The treatments are 10 varieties of common beans.

In Phase I, these are grown in a field, in 10 blocks of size 6.

In Phase II, a sample of beans is taken from each plot.
Each sample is cooked in a special machine. The measured response is the time taken to properly cook the beans.

In Phase II, only four samples can be processed per day.
So we should treat days as 15 blocks of size 4.

Example 2: A two-phase experiment in agriculture

The treatments are 10 varieties of common beans.

In Phase I, these are grown in a field, in 10 blocks of size 6.

In Phase II, a sample of beans is taken from each plot.
Each sample is cooked in a special machine. The measured response is the time taken to properly cook the beans.

In Phase II, only four samples can be processed per day.
So we should treat days as 15 blocks of size 4.

Now the design consists of one function allocating bean varieties to plots in the field, and another function allocating each plot to a run of the cooking machine.

Model when there are two systems of blocks

We measure the response Y on each sample.

If that sample is from a plot in block m with treatment i in Phase I and it is allocated to day n in Phase II, then we assume that

$$Y = \tau_i + \beta_m + \gamma_n + \text{random noise.}$$

To get rid of the β parameters and the γ parameters, we look at $(I - P_*)Y$, where P_* is the $N \times N$ matrix of orthogonal projection onto the space spanned by the characteristic vectors of the blocks in Phase I and the characteristic vectors of the days in Phase II.

Model when there are two systems of blocks

We measure the response Y on each sample.

If that sample is from a plot in block m with treatment i in Phase I and it is allocated to day n in Phase II, then we assume that

$$Y = \tau_i + \beta_m + \gamma_n + \text{random noise.}$$

To get rid of the β parameters and the γ parameters, we look at $(I - P_*)Y$, where P_* is the $N \times N$ matrix of orthogonal projection onto the space spanned by the characteristic vectors of the blocks in Phase I and the characteristic vectors of the days in Phase II.

Let X be the $N \times v$ incidence matrix of treatments in experimental units.
The information matrix is $X^\top (I - P_*)X$.

Computer search

At a conference on variety-testing in Słupia Wielka, Poland, in June 2018, Nha Vo-Thanh (Universität Hohenheim) gave a talk explaining his work with Hans-Peter Piepho on several different methods of computer search to find a good design for this experiment.

Computer search

At a conference on variety-testing in Słupia Wielka, Poland, in June 2018, Nha Vo-Thanh (Universität Hohenheim) gave a talk explaining his work with Hans-Peter Piepho on several different methods of computer search to find a good design for this experiment.

That evening, I got out some paper and a pen, and scribbled down some ideas, using my combinatorial approach. Very soon, I had a design with a smaller value of $\bar{V}$ than he had found.

Computer search

At a conference on variety-testing in Słupia Wielka, Poland, in June 2018, Nha Vo-Thanh (Universität Hohenheim) gave a talk explaining his work with Hans-Peter Piepho on several different methods of computer search to find a good design for this experiment.

That evening, I got out some paper and a pen, and scribbled down some ideas, using my combinatorial approach. Very soon, I had a design with a smaller value of $\bar{V}$ than he had found.

Here $\bar{V}$ denotes the average of the variances of the estimators of the differences between pairs of different treatments.
It can be calculated from the information matrix.

Principle: Consider the smaller blocks first

The blocks in Phase II are smaller than those in Phase I,
so they will have more effect on increasing the variance.
So it makes sense to consider the design for Phase II first.

Principle: Consider the smaller blocks first

The blocks in Phase II are smaller than those in Phase I,
so they will have more effect on increasing the variance.
So it makes sense to consider the design for Phase II first.

There are 10 treatments in 15 blocks of size 4.
Think of the treatments as all pairs from $\{1, 2, 3, 4, 5\}$.
An obvious way to make 15 blocks of size 4 is to use the
4-cycles in the complete graph K_5 on 5 vertices.

Principle: Consider the smaller blocks first

The blocks in Phase II are smaller than those in Phase I,
so they will have more effect on increasing the variance.
So it makes sense to consider the design for Phase II first.

There are 10 treatments in 15 blocks of size 4.
Think of the treatments as all pairs from $\{1, 2, 3, 4, 5\}$.
An obvious way to make 15 blocks of size 4 is to use the
4-cycles in the complete graph K_5 on 5 vertices.

In fact, this design is balanced (all concurrences are 2),
so it is best possible for Phase II.

Principle: Consider the smaller blocks first

The blocks in Phase II are smaller than those in Phase I,
so they will have more effect on increasing the variance.
So it makes sense to consider the design for Phase II first.

There are 10 treatments in 15 blocks of size 4.
Think of the treatments as all pairs from $\{1, 2, 3, 4, 5\}$.
An obvious way to make 15 blocks of size 4 is to use the
4-cycles in the complete graph K_5 on 5 vertices.

In fact, this design is balanced (all concurrences are 2),
so it is best possible for Phase II.

A	B	C	E	F	G	H	D	B	D	A	F	A	E	B
E	H	F	H	J	I	J	I	C	J	C	G	D	G	I
C	A	B	G	E	F	B	C	I	A	G	C	E	B	A
H	F	E	J	I	H	D	H	J	F	J	D	I	D	G

Principle: Consider the smaller blocks first

The blocks in Phase II are smaller than those in Phase I, so they will have more effect on increasing the variance. So it makes sense to consider the design for Phase II first.

There are 10 treatments in 15 blocks of size 4.
Think of the treatments as all pairs from $\{1, 2, 3, 4, 5\}$.
An obvious way to make 15 blocks of size 4 is to use the 4-cycles in the complete graph K_5 on 5 vertices.

In fact, this design is balanced (all concurrences are 2), so it is best possible for Phase II.

A	B	C	E	F	G	H	D	B	D	A	F	A	E	B
E	H	F	H	J	I	J	I	C	J	C	G	D	G	I
C	A	B	G	E	F	B	C	I	A	G	C	E	B	A
H	F	E	J	I	H	D	H	J	F	J	D	I	D	G

Principle: Consider the smaller blocks first

The blocks in Phase II are smaller than those in Phase I,
so they will have more effect on increasing the variance.
So it makes sense to consider the design for Phase II first.

There are 10 treatments in 15 blocks of size 4.
Think of the treatments as all pairs from $\{1, 2, 3, 4, 5\}$.
An obvious way to make 15 blocks of size 4 is to use the
4-cycles in the complete graph K_5 on 5 vertices.

In fact, this design is balanced (all concurrences are 2),
so it is best possible for Phase II.

A	B	C	E	F	G	H	D	B	D	A	F	A	E	B
E	H	F	H	J	I	J	I	C	J	C	G	D	G	I
C	A	B	G	E	F	B	C	I	A	G	C	E	B	A
H	F	E	J	I	H	D	H	J	F	J	D	I	D	G

The non-intuitive step

The Phase II design has the property that we can group its days into five groups of three days, in such a way that every treatment in a group occurs twice in that group.

The non-intuitive step

The Phase II design has the property that we can group its days into five groups of three days, in such a way that every treatment in a group occurs twice in that group.

Arrange each group as a $(2 \times 3)/2$ rectangle,
in such a way that days are columns
and each treatment in the group occurs in both rows.

The non-intuitive step

The Phase II design has the property that we can group its days into five groups of three days, in such a way that every treatment in a group occurs twice in that group.

Arrange each group as a $(2 \times 3)/2$ rectangle,
in such a way that days are columns
and each treatment in the group occurs in both rows.

A	B	C	E	F	G	H	D	B	D	A	F	A	E	B
E	H	F	H	J	I	J	I	C	J	C	G	D	G	I
C	A	B	G	E	F	B	C	I	A	G	C	E	B	A
H	F	E	J	I	H	D	H	J	F	J	D	I	D	G

The non-intuitive step

The Phase II design has the property that we can group its days into five groups of three days, in such a way that every treatment in a group occurs twice in that group.

Arrange each group as a $(2 \times 3)/2$ rectangle,
in such a way that days are columns
and each treatment in the group occurs in both rows.

A	B	C	E	F	G	H	D	B	D	A	F	A	E	B
E	H	F	H	J	I	J	I	C	J	C	G	D	G	I
C	A	B	G	E	F	B	C	I	A	G	C	E	B	A
H	F	E	J	I	H	D	H	J	F	J	D	I	D	G

Use each row as a field block in Phase I.

The non-intuitive step

The Phase II design has the property that we can group its days into five groups of three days, in such a way that every treatment in a group occurs twice in that group.

Arrange each group as a $(2 \times 3)/2$ rectangle,
in such a way that days are columns
and each treatment in the group occurs in both rows.

A	B	C	E	F	G	H	D	B	D	A	F	A	E	B
E	H	F	H	J	I	J	I	C	J	C	G	D	G	I
C	A	B	G	E	F	B	C	I	A	G	C	E	B	A
H	F	E	J	I	H	D	H	J	F	J	D	I	D	G

Use each row as a field block in Phase I.
The treatment information lost to field blocks is
the same as the information lost to rectangles,
which is part of the information already lost to days,
so no further information is lost in Phase I.

A surprising theorem

Theorem

In a nested row-column design,
if the rows within each rectangle have exactly the same treatments
then the loss of information on treatment differences is the same
as it is in the block design obtained by ignoring rectangles and rows.

A surprising theorem

Theorem

In a nested row-column design,
if the rows within each rectangle have exactly the same treatments
then the loss of information on treatment differences is the same
as it is in the block design obtained by ignoring rectangles and rows.

This property is known as adjusted orthogonality. In this case, treatments and Phase I blocks have adjusted orthogonality with respect to Phase II blocks because the corresonding vector spaces are orthogonal to each other after we have projected them both orthgonally to Phase II blocks.

A surprising theorem

Theorem

In a nested row-column design,
if the rows within each rectangle have exactly the same treatments
then the loss of information on treatment differences is the same
as it is in the block design obtained by ignoring rectangles and rows.

This property is known as adjusted orthogonality. In this case, treatments and Phase I blocks have adjusted orthogonality with respect to Phase II blocks because the corresonding vector spaces are orthogonal to each other after we have projected them both orthgonally to Phase II blocks.

In this example, the best design for Phase I alone cannot be arranged as a nested row-column design with this property.

Comparison of designs

Designs are often compared by using the A criterion.
This is the inverse of $\bar{V}$,
scaled to have value 1 for an unblocked equireplicate design.

Comparison of designs

Designs are often compared by using the A criterion.
This is the inverse of $\bar{V}$,
scaled to have value 1 for an unblocked equireplicate design.

Design	computer search	patterns
A	0.80896	0.83333

Comparison of designs

Designs are often compared by using the A criterion.
This is the inverse of $\bar{V}$,
scaled to have value 1 for an unblocked equireplicate design.

Design	computer search	patterns
A	0.80896	0.83333

After I showed my results to NVT and HPP,
they adapted their search method to incorporate that theorem.